A Focused Information Criterion for Graphical Models

Eugen Pircalabelu · Gerda Claeskens · Lourens Waldorp

May 8, 2014

Abstract A new method for model selection for Gaussian Bayesian networks and Markov networks, with extensions towards ancestral graphs, is constructed to have good mean squared error properties. The method is based on the focused information criterion, and offers the possibility of fitting individual-tailored models. The focus of the research, that is, the purpose of the model, directs the selection. It is shown that using the focused information criterion leads to a graph with small mean squared error. The low mean squared error ensures accurate estimation using a graphical model; here estimation rather than explanation is the main objective. Two situations that commonly occur in practice are treated: a data-driven estimation of a graphical model and the improvement of an already pre-specified feasible model. The search algorithms are illustrated by means of data examples and are compared with existing methods in a simulation study.

Keywords Focused information criterion · Model selection · Gaussian Bayesian network · Gaussian Markov network · Directed acyclic graph · Ancestral graph

1 Introduction

Probabilistic graphical models are increasingly studied in the statistical and machine learning community, because of their direct applicability to problems coming from areas such as image analysis, engineering, biomedical and computer sciences. The most popular such graphical structures are the Bayesian networks and Markov networks which have been extensively studied in Whittaker (1990), Lauritzen (1996), Edwards (2000), Koller and Friedman (2009), amongst others.

The main objective of the present paper is to develop a methodology to estimate a directed graphical model such that the selected model has good performance with respect to estimation of a set of model parameters. For each node in the graph we define a focus parameter and estimate a nodewise model. For instance, we could be interested in estimating the score of a student's performance on algebra or any other subject and as a consequence of our procedure, estimation or prediction of such scores is based on nodewise models. We proceed by combining all nodewise models to construct a global graph estimate. In

E. Pircalabelu, G. Claeskens
ORSTAT and Leuven Statistics Research Center
KU Leuven
Naamsestraat 69, 3000 Leuven, Belgium
E-mail: eugen.pircalabelu@kuleuven.be
E-mail: gerda.claeskens@kuleuven.be

L. Waldorp
Department of Psychological Methods
University of Amsterdam
Weesperplein 4, 1018 Amsterdam, The Netherlands
E-mail: waldorp@uva.nl

order for this estimation to be as accurate as possible, we construct our model to have low mean squared error for this estimate (Claeskens and Hjort, 2008b; Hastie et al, 2009). The emphasis of the score that will be used is therefore on estimation and not on explanation. We concentrate in this paper on small and moderately sized networks, as the treatment of larger networks is kept for a separate study since these require alternative, penalized, methods and possibly different search methods than what we currently use. The methods proposed here deal with the 'regular' case where the number of cases n is comfortably larger than the number of unknown parameters to estimate.

For this purpose we use the focused information criterion, FIC (Claeskens and Hjort, 2003). Unlike other information criteria, such as the traditional Akaike information criterion (AIC, Akaike, 1973) and the Bayesian information criterion (BIC, Schwarz, 1978), the FIC allows for selecting individual models, tailored to a specific purpose (the focus), as opposed to attempting an identification of a single model that should be used for all purposes. The AIC/BIC approach ignores the possibility of using different models for different estimation problems, as in estimating different scores of students, say. The FIC minimizes the mean squared error (MSE) for the focus estimator, which is a sum of squared bias and variance.

The main difference between working with the FIC as opposed to working with penalized likelihood criteria, is that with FIC, models are selected based directly on their MSE performance. This seems to be a fruitful line of thought, as one normally would be interested in estimators with low MSE, as this balances the main quantities of interest: bias and variance of the estimator. Additionally, different specifications for the focus parameter of interest, denoted as μ, may lead to ultimately different models being selected because some models will perform better for some focuses and worse for others. For instance, a model to predict a student's score of algebra could be different than a model to predict a student's score of analysis. The choice of the focus μ reflects different and possibly divergent research objectives. For example, one researcher might be interested in expected values at fixed positions in the covariate space, while another might be interested only in the impact of a particular covariate on the mean structure. A third researcher might be interested in the 90th quantile of the response distributions at a fixed covariate position, and a fourth researcher might want to estimate the generalized variance, that is the determinant of the covariance matrix of the data. This starting point of the model selection procedure should be contrasted with using a BIC/AIC criterion which does not include such modeling aspects and outputs always a single model, namely the best penalized likelihood model, that is supposed to be used for all purposes. The penalty in the 'penalized likelihood' class of criteria is needed since the maximized likelihood always increases when more parameters are added to the model. Without a penalty, the largest model would always get selected. Adding a penalty provides a trade-off between a good fit and the complexity of the model. A similar phenomenon appears for the FIC. The wide model (or full model), the most complex model one is willing to consider to estimate μ produces the smallest bias, but the largest variance. The simplest model in the search list, the narrow model, contains the fewest parameters to estimate and hence comes with a possible large bias, though a small variance. Both models are likely to result in large MSE values. A model somewhere in between the narrow and full model might better balance the contributions of squared bias and variance, thus producing lower MSE values. This is exactly what the FIC does: obtain low MSE values by estimating the MSE for a specific focus. Models with a low MSE to estimate μ lead to a more accurate estimation.

The proposed search algorithms fall into the category of 'score-based' model selection, where models receive a corresponding score, in this case based on an estimated mean squared error of the focus estimator. While the FIC has been applied to generalized linear models (Claeskens and Hjort, 2008a), in models for survival data (Hjort and Claeskens, 2006), in generalized partial linear models (Zhang and Liang, 2011), and in several other types of models, its definition, computation and application to graphical models is new. The advantages of using an FIC estimated model are two-fold: first, the estimated graph will provide low MSE for the selected focus estimator even without the assumption that the model is correct (see Section 4), and second, models can be selected at different levels of analysis, e.g., given a particular configuration for an individual, a model for such a configuration can be selected. Selecting one 'average' model, though still having low MSE for the focus estimator, is another possibility of the focused selection.

A focused model search comes close to the goals of the concept of 'personalized medicine' (see Shastry, 2006; Mansour and Schwarz, 2008; van 't Veer and Bernards, 2008) that seems to be increasingly more embraced by practitioners, where different actions might be taken for each subject based on his/her personal

characteristics and predispositions. Subject-specific models like the ones estimated with FIC, might aid for such an ambitious goal.

2 Focused model search, a simple example

In this example we show how the FIC can be used to select a subset of edges to estimate a plausible structure of a small network. For this example we use the 'Mathematics grades' dataset (Mardia et al, 1979) which contains $n = 88$ observations for each 5 variables: Mechanics (MEC), Vectors (VEC), Algebra (ALG), Analysis (ANL) and Statistics (STA).

We start by presenting how the FIC score is used at the level of node, and without loss of generality assume the node under consideration is ALG. On the standardized dataset, for the variable ALG, we define the parameter of interest, or focus, as the conditional mean for specific scores on the covariates $z = (\text{MEC, VEC, ANL, STA})$, i.e., $\mu = E[\text{ALG}|z = (2.18, 2.39, 1.37, 2.24)]$. In words, we wish to estimate the expected value of that student's algebra score, given the results on the other four examination scores. We then fit the normal linear regression model using all of the 88 observations,

$$\text{ALG}_k = \beta_0 + \beta_1\text{MEC}_k + \beta_2\text{VEC}_k + \beta_3\text{ANL}_k + \beta_4\text{STA}_k + \epsilon_k, \quad k = 1, \ldots, n. \tag{1}$$

The FIC is now used to select which coefficients are recommended to estimate the conditional mean. A subscript S indicates the (sub)model that is used. Let $\hat{\mu}_S$ denote the maximum likelihood estimator (MLE) of μ when model S is considered. For example, with the choice $S = \{1, 2\}$, $\hat{\mu}_S$ denotes the MLE of μ when only the covariates MEC and VEC are included in the model. Each such choice of S gives rise to an estimator with its own bias (not all relevant variables might have been included) and variance (some redundant variables might lead to an increased variance). The FIC compares the estimated mean squared error of $\hat{\mu}_S$ for different possible submodels S and selects the model with the smallest MSE.

In the above argumentation we have concentrated on only one node. In order to estimate a graph our approach uses a hill-climbing procedure (see Section 4.2 for the algorithm) to estimate a global DAG using nodewise model selection. We describe below the different steps how one can estimate the DAG for the top ranked student as presented in Figure 2.

The starting point is the vector of standardized grades which corresponds to the subject of interest, that is (MEC, VEC, ALG, ANL, STA)=(2.18, 2.39, 1.54, 1.37, 2.24) after which we compute the FIC scores at each node in the empty graph, where none of the nodes has any parent and the conditional expectation is the focus parameter, resulting in $(\text{FIC}_{\text{MEC}}, \text{FIC}_{\text{VEC}}, \text{FIC}_{\text{ALG}}, \text{FIC}_{\text{ANL}}, \text{FIC}_{\text{STA}}) = (144.9, 148.8, 377.5, 195.9, 97.1)$. The total FIC score for the initial empty graph is the sum of the nodewise values (i.e. 964.2). This value is expected to be quite large, we expect the estimated bias to be large for the empty graph.

The greedy search algorithm needs for this example 9 iterations until it converges; Figure 1 presents step-by-step how the final DAG was obtained. We now describe the first iteration. Starting from the empty graph there are 20 possible DAGs containing only one edge. These 20 one-edge DAGs can all be scored in parallel. We start for example with the DAG having only the connection MEC$\rightarrow$VEC (the rest of the graph remains empty at this stage). We take VEC as the node on which we concentrate, the focus, since this is a child node. The estimated MSE of this one-edge model for VEC turns out to be FIC_{VEC}=4.49, while the new nodewise FIC score vector becomes $(\text{FIC}_{\text{MEC}}, \text{FIC}_{\text{VEC}}, \text{FIC}_{\text{ALG}}, \text{FIC}_{\text{ANL}}, \text{FIC}_{\text{STA}}) = (144.9, 4.5, 377.5, 195.9, 97.1)$, with a total FIC for the current DAG equal to 819.86. Although this graph scores lower than the empty graph, we can still do better. Out of all 20 DAGs, the lowest total graph FIC value is obtained for the DAG containing STA $\rightarrow$ ALG with $(\text{FIC}_{\text{MEC}}, \text{FIC}_{\text{VEC}}, \text{FIC}_{\text{ALG}}, \text{FIC}_{\text{ANL}}, \text{FIC}_{\text{STA}})$ =(144.9, 148.8, 30.8, 195.9, 97.1) and for which the total score is equal to 617.5. We accept this move in the algorithm, since it is the best scoring move, update the empty graph by including the edge STA $\rightarrow$ ALG and proceed forward.

In the second iteration, we start from the DAG that resulted from iteration 1, and construct new graphs based on all possible moves that involve modifying it by just one edge. There are again 20 possible moves that result in a valid DAG specification (18 single edge additions, deleting the edge introduced at iteration 1 or reversing it). All the DAGs are scored in parallel, and for example, one such DAG contains the edges MEC $\rightarrow$ ALG $\leftarrow$ STA and thus, the focus is now the conditional expectation of ALG given MEC and STA. After evaluating all 20 possible moves at iteration 2, each time taking a child node as a focus node in the

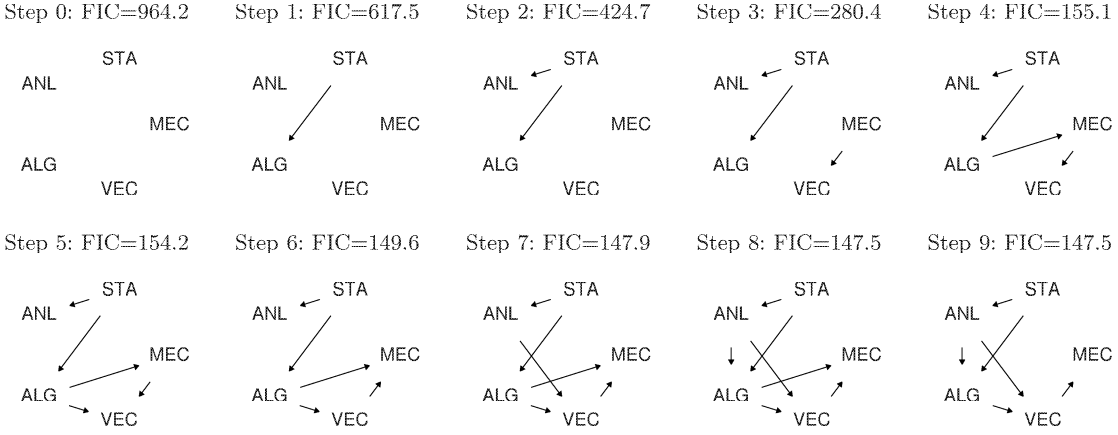


Fig. 1: Mathematics grades. Iterative steps for estimating a DAG by FIC for the top ranked student.

nodewise FIC value, amongst these 20 DAGs, the best total FIC value is obtained when introducing in the DAG also the edge STA→ ANL leading to a new FIC score of 424.7. The same process is repeated until no improvement in the total FIC score is possible or when all possible moves would result in violating the DAG constraints.

Using the focused selection procedure as explained intuitively above, we estimate for the top student and the bottom ranked student the two DAGs presented in Figure 2. The graph for the bottom ranked student is obtained in the same manner as for the top ranked student, the only difference being that we start from a different vector of standardized grades, i.e. (MEC, VEC, ALG, ANL, STA)=$(-2.23, -0.81, -2.79, -2.54, -1.64)$, to use as a covariate vector for computing the FIC values. When comparing the two estimated networks, it is apparent that relations such as ANL $\rightarrow$ ALG $\rightarrow$ VEC are common, while the remaining ones are quite different between the two subjects, either reversed such as in the relation connecting ALG and STA, or present in only one of the graphs such as STA $\rightarrow$ ANL.

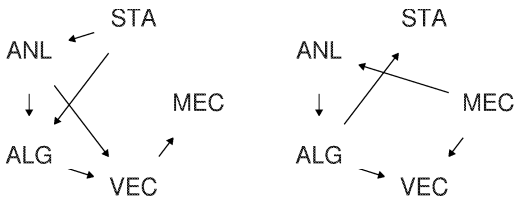


Fig. 2: Mathematics grades. Selected DAGs using the FIC for (a) the top ranked student and (b) the bottom ranked student.

3 Notation for Graphical Models

3.1 Graphical models

We denote by $\mathcal{G}(\mathcal{E}, \mathcal{V})$ a graphical structure (or graph) based on a set of nodes ($\mathcal{V}$), a set of edges ($\mathcal{E}$), and a set of random variables $\{X_i : i \in \mathcal{V}\}$. For $\mathcal{V} = \{1, \ldots, p\}$, each of the variables $X_1, \ldots, X_p$ corresponds to one node (or vertex) in the set $\mathcal{V}$. The set of edges (or arcs) $\mathcal{E}$ is a subset of $\mathcal{V} \times \mathcal{V}$, the set of ordered pairs of distinct nodes. A connection between two nodes (say, i and j) can be either undirected ($i - j$) or directed ($i \to j$ or $i \leftarrow j$). We denote a directed edge $j \leftarrow i$ in $\mathcal{E}$ by (j, i) and call node i (or variable X_i) a *parent* of node j (or variable X_j), conversely node j is referred to as a *child* of node i. To make the notation easier, an undirected edge $i - j$ is set between nodes i and j if and only if $\mathcal{E}$ contains both (i, j) and (j, i), and call i and j adjacent (or *neighbors*). A directed path between nodes i and k is a sequence of nodes that starts in i and by following the directionality of the arrows leads to node k (e.g. $i \to j \to \ldots \to y \to k$). Node i is referred to as an *ancestor* of k if there exists such a directed path between the two nodes, or if $i = k$. Similarly k is then said to be a *descendant* of i. For later use, we define a third type of arrow $i \leftrightarrow j$ which will be used to refer two nodes as being 'spouses'. Only one connection can be made between two nodes i and j, be it directed, undirected or bidirected, and no self-loop edges are permitted (such as $i \to i$, $i - i$ or $i \leftrightarrow i$).

3.2 Directed Acyclic Graphs and Bayesian Networks

We consider Bayesian networks (BN) as a class of statistical models, consisting of a graph $\mathcal{G}(\mathcal{E}, \mathcal{V})$ and a probability distribution f, with two particular characteristics. First, the graph $\mathcal{G}(\mathcal{E}, \mathcal{V})$ contains only directed edges between pairs of vertices, such that there are no feedback loops (referred to as the 'acyclicity' property). That is, any directed path starting at node i cannot lead back to i. Because of this property, such a graph is called a 'directed acyclic graph' (DAG). Second, the joint multivariate probability density function (pdf) of $(X_1, \ldots, X_p)$ factorizes as

$$f(x_1, \ldots, x_p) = \prod_{l=1}^{p} f(x_l | pa(x_l)),$$

where the conditioning is on $pa(x_l)$, the set of parental variables of X_l (see Lauritzen, 1996). Graphically, this is represented by a directed arrow from each of the 'parents' to the 'children'.

We further say that f has the local Markov property with respect to $\mathcal{G}$ if

$$\forall l \in \mathcal{V}, \quad l \perp nd(l) | pa(l)$$

where the symbol $\perp$ denotes independence and $nd(l)$ denotes the set of non-descendants of node l. That is, any node is independent of its non-descendants when conditioned on its corresponding parents. Moreover, if f admits a factorization according to $\mathcal{G}$ then $\mathcal{A} \perp \mathcal{B} | \mathcal{C}$ if for any $\mathcal{A}, \mathcal{B}, \mathcal{C}$ disjoint subsets of $\mathcal{V}$ where $\mathcal{C}$ separates $\mathcal{A}$ and $\mathcal{B}$ in $\mathcal{G}^m_{an(\mathcal{A} \cup \mathcal{B} \cup \mathcal{C})}$, which is a moralized graph containing the ancestral set of $\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$. By moralization we mean the process where 'unmarried' parents having a common child get connected (or 'married') by an undirected link and all arrows get dropped. The moralized graph is thus an undirected one.

In the Gaussian Bayesian net all conditional pdfs are linear Gaussians (see Koller and Friedman, 2009, chap. 5 & 7), where X_l has a linear Gaussian model if conditional on its parents, $pa(X_l)$,

$$X_l | pa(X_l) \sim N(\beta_{l0} + \beta_{l1} x_1 + \ldots + \beta_{li} x_i; \sigma^2), \tag{2}$$

for some values of the regression coefficients $\beta_{l0}, \ldots, \beta_{li}$ and of the variance σ^2.

3.3 Markov networks

If all edges in $\mathcal{E}$ are undirected, we call $\mathcal{G}(\mathcal{E}, \mathcal{V})$ an undirected graph or a Markov network (MN), to which (as in the BN case) a pairwise, a local and a global Markov property can be associated. A probability distribution is said to have a pairwise Markov property relative to $\mathcal{G}$ if for two non-adjacent nodes i and j, $i \perp j | \mathcal{V} \setminus \{i, j\}$ and a global Markov property if for any $\mathcal{A}, \mathcal{B}, \mathcal{C}$ disjoint subsets of $\mathcal{V}$ where $\mathcal{C}$ separates $\mathcal{A}$ and $\mathcal{B}$ in graph $\mathcal{G}$, we have $\mathcal{A} \perp \mathcal{B} | \mathcal{C}$ (Lauritzen, 1996).

If all conditional independencies that can be read from the graph and all the conditional independencies that hold in the distribution are equivalent, then we call graph $\mathcal{G}$ a 'perfect map' of the distribution in the terminology of Pearl (1988), or call them 'faithful to one another' in the terminology of Spirtes et al (2000).

We assume that $(X_1, \ldots, X_p) \sim N_p(\mu, \Sigma)$, where without loss of generality we take $\mu = 0$, and in such a case the global and local Markov properties coincide. Proposition 5.2 in Lauritzen (1996) asserts that if $(X_1, \ldots, X_p) \sim N_p(0, \Sigma)$, then

$$X_i \not\perp X_j | X_{\mathcal{V} \setminus \{i,j\}} \Leftrightarrow \Sigma^{-1}_{ij} \neq 0$$

Moreover, $\Sigma^{-1}_{ij} \neq 0 \Leftrightarrow \beta_{ij} \neq 0 \ \& \ \beta_{ji} \neq 0$ where the following conditional equation for any particular variable is put forward,

$$X_j | \{X_i : i \in \mathcal{V} \setminus j\} = \sum_{i \in \mathcal{V} \setminus j} \beta_{ji} X_i + \epsilon_j, \quad \epsilon_j \sim N(0, \Sigma_{jj}), \quad \forall j \in \mathcal{V}.$$

Thus using the Markov property and assuming faithfulness as well as a multivariate normal distribution, one can conclude that an undirected arrow exists between X_j and X_i, if and only if there are non-zero partial regression coefficients in both of the following regression models: regressing X_i on X_j (given all other nodes) and regressing X_j on X_i (given all other nodes).

The concentration matrix Σ^{-1} contains information about the covariance between pairs (X_i, X_j) conditioned on all other variables in the model. This motivates the names 'concentration graph' and 'covariance selection model' (Dempster, 1972).

The implication of the above reasoning is that the entire covariance estimation problem can be translated into regression language (see also Meinshausen and Bühlmann, 2006), which will be exploited in the focused model search.

3.4 Ancestral Graphs

Richardson and Spirtes (2002) introduced the 'ancestral graphs' (AG) as a specification of a general 'mixed graph' that can contain three types of edges: undirected ($i - j$), directed ($i \rightarrow j$ or $i \leftarrow j$) and bidirected ($i \leftrightarrow j$). The advantage of this class of models is that ancestral graphs are closed under both conditioning and marginalization. Not all configurations of edges are allowed in an ancestral graph, which is a graph constrained by two conditions specifying that there can be no cycles and no directed edges (coming from parents or spouses) to nodes with undirected edges, i.e., for all nodes i:

$$i \notin an(pa(i) \cup sp(i)) \text{ and if } ne(i) \neq \emptyset \text{ then } pa(i) \cup sp(i) = \emptyset,$$

where $an(i), pa(i), sp(i), ne(i)$ are the corresponding sets of ancestors, parents, spouses and neighbors of node i. The first condition just translates to saying that if an edge such as $i \leftarrow j$ exists in the graph, then i cannot be an ancestor of j.

A more intuitive definition of what an ancestral graph can contain is offered in Ali et al (2009). A graph $\mathcal{G}$ is ancestral if it does not contain any directed cycles, and if an edge such as $i \leftrightarrow j$ appears, then there is no directed path from i to j (or j to i), while if the edge $i - j$ appears, then both i and j should not have spouses or parents.

Under the assumption of joint multivariate normality for the variables of interest, Drton and Richardson (2004) have provided an algorithm to estimate the covariance matrix Σ, using the idea of *iterative conditional fitting* (ICF), where the matrix Σ is decomposed as:

$$\Sigma = (I - B)^{-1} \begin{pmatrix} \Lambda^{-1} & 0 \\ 0 & \Omega \end{pmatrix} \left((I - B)^{-1}\right)^T .$$

The matrices B, Λ, and Ω contain non-zero entries corresponding respectively, to the set of directed ($B = (\beta_{ji})$), undirected ($\Lambda = (\lambda_{ji})$) and bidirected ($\Omega = (\omega_{ji})$) edges. The interpretation of the entries in the three matrices is as follows: λ_{ji} represents an inverse covariance element for the subgraph formed by the undirected edges, $-\beta_{ji}$ represents the partial regression coefficient of node i in the relation $j \leftarrow i$ and ω_{ji} represents the covariance between errors ϵ_i and ϵ_j once the influence of both sets of parents $pa(i)$ and $pa(j)$ has been controlled for, in the regression model $X_i = \sum_{l \in pa(i)} \beta_{il} X_l + \epsilon_i$ and similarly for X_j.

Note that the ICF algorithm only performs estimation, rather than estimating the structure of the graph (i.e. for a given user prespecified graphical structure, the algorithm estimates the B, Λ, Ω matrices). Spirtes et al (1999) and Zhang (2008) have proposed a constraint-based approach which using a series of orientation rules, estimates from the data a partial ancestral graph (PAG), that describes the Markov equivalence class of an AG. It is a partial AG, in the sense that, there are possibly six kinds of edges $-, \rightarrow, \leftrightarrow, \circ\!\!-, \circ\!\!-\!\!\circ, \circ\!\!\rightarrow$, as the extra $\circ$ symbol denotes an undetermined edge mark.

4 FIC for Model Selection in Graphical Models

4.1 Estimating the mean squared error

Consider a dataset consisting of n independent cases for each p-dimensional vector $(X_{k1}, \ldots, X_{kp})$ where for each $k = 1, \ldots, n$ we assume a linear Gaussian model such that for each fixed $i \in \mathcal{V}$, the variables $\{X_{ki:k=1,\ldots,n}\}$ are independent. That is, conditional on $\{X_{ki} : i \in \mathcal{V} \setminus j\}$

$$X_{kj} | \{X_{ki} : i \in \mathcal{V} \setminus j\} \sim N\left(\sum_{i \in \mathcal{V} \setminus j} \beta_{ji} X_{ki}, \Sigma_{jj} \right) \ , \ \forall j \in \mathcal{V} \ \& \ \forall k = 1, \ldots, n.$$

We define the vector θ_j to contain the parameters that should be estimated in all models and that are never subject to model selection or exclusion. For example, regardless of which parents enter the model, Σ_{jj} has to be estimated in all considered models. Hence Σ_{jj} is one of the components of θ_j. If based on theoretical reasons, a variable X_q is decided beforehand to be a parent of X_j regardless of what other variables are selected as parents, then also β_{jq} is included in θ_j.

To facilitate model selection, for each $j \in \mathcal{V}$ we introduce the vector (of length $p - 1$) γ_j with ith element ($i \in \mathcal{V}$)

$$\gamma_{ji} = \begin{cases} \beta_{ji} & \text{if } X_i \text{ is a parent of } X_j \\ 0 & \text{otherwise.} \end{cases}$$

The vector γ_j is distinct from θ_j, model parameters are either included in θ_j or in γ_j, never in both vectors.

For each node $j \in \mathcal{V}$, a subset $S \subseteq \mathcal{V} \setminus j$ of possible parents is to be selected. To simplify the notation, we omit the index j since in the remainder of this section all derivations are nodewise. Later, when necessary, the subscript j is reintroduced. There is a one-to-one correspondence between S and γ. For example, the largest such subset, denoted as S_{wide} corresponds to setting the γ parameters to non-zero values, while an empty set S corresponds to all the γ parameters set to zero and thus, indicating no parents. For any other S in between the wide and empty set, particular elements of the γ vector are set to 0, while others are not. The length of γ is always $p - 1$. When interest is in the part of γ that contains non-zero elements as dictated by using a submodel indexed by S, we denote by γ_S the subvector of γ formed by components γ_{ji} for which $i \in S$. Each model based on S corresponds to working with the density $f(X; \theta, \gamma_S)$ where particular parameters in the γ vector have been set to 0, according to S (Claeskens and Hjort, 2008b).

Most often, the list of possible parents of variable X is left unspecified, such that an extensive search could be performed. In other situations this list can be constrained beforehand, if knowledge about plausible or implausible relations is available to the researcher.

For any particular node, we define a *focus parameter* $\mu = \mu(\theta, \gamma; x)$ that is a function of the θ and γ parameters of the density, and potentially of a user-specified vector of covariate values x, allowing for personalized model search. The focus is intended to be estimated as precisely (in the MSE sense) as possible. We assume that μ is differentiable with respect to θ and γ, and in each model based on S, a maximum likelihood estimator $\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S; x)$ is constructed. The FIC method estimates $\text{MSE}(\hat{\mu}_S)$ and selects the model with the smallest such value. Note that the length of the vector $\hat{\theta}_S$ is always the same, its value, though, may change with different choices of S.

Thus, for different focuses different orderings of the MSE values might occur, leading to possibly different selected models, depending on the specific focus or target. In this way, one can obtain better selected models in terms of MSE than obtained from a global model search not taking any use of the selected model (focus) into account. For this particular application of FIC for estimating graphs, the focus is the expected value of a variable, reflecting the interest in discovering a topology of the graph that produces a low MSE of the expected value. For examples of other focuses, see Section 6.2.

We further introduce the Fisher information matrix for model S, i.e., the expected value of the matrix of minus second partial derivatives of the log-likelihood with respect to the parameters, $J_S = \begin{pmatrix} J_{00,S} & J_{01,S} \\ J_{10,S} & J_{11,S} \end{pmatrix}$, partitioned in blocks according to the length of θ and the number of elements in S. We define $Q_S = J^{-1}_{11,S}$, $Q = J^{-1}_{11}$, with J without subscript being the Fisher information matrix for the 'wide' model, $Q^0_S = \pi^t_S Q_S \pi_S$ the matrix with the same dimensions as Q and with elements equal to those of Q_S for those rows and columns indexed by elements from S, and all other elements zero, and $G_S = \pi^t_S Q_S \pi_S Q^{-1}$. The projection matrix π_S with dimension $|S| \times (p-1)$, that has been used to define Q^0_S and G_S, contains 0s and 1s, such that when multiplied with matrices of interest, it retains those rows and columns that relate to the parameters contained in model S, e.g. for a vector v, with $\pi_{\{2\}} = (0, 1, 0, \ldots, 0)$, $\pi_{\{2\}} v = v_2$, the second component of v (see Claeskens and Hjort, 2008b, p. 146).

To balance the contributions of the squared bias and variance of the estimators $\hat{\mu}_S$, similarly as in Hjort and Claeskens (2003), we consider a local misspecification setting, that is, each X_{kj}, for $k = 1, \ldots, n$, has pdf

$$f(x_j | pa(x_j); \theta_0, \gamma_0 + \delta/\sqrt{n}),$$

where θ_0 and γ_0 correspond to the narrow model where S is the empty set and γ_0 is a vector of zeros only. Theorem 6.1 in Claeskens and Hjort (2008b) asserts that under certain conditions the maximum likelihood estimator of the focus parameter obeys

$$\sqrt{n}\,(\hat{\mu}_S - \mu_{true}) \xrightarrow{d} \Lambda_0 + \omega^t(\delta - G_S D),$$

where $\Lambda_0 \sim N\left(0, \left(\frac{\partial \mu}{\partial \theta}\right)^t J^{-1}_{00} \frac{\partial \mu}{\partial \theta}\right)$, $\omega = J_{10} J^{-1}_{00} \frac{\partial \mu}{\partial \theta} - \frac{\partial \mu}{\partial \gamma}$, $D \sim N(\delta, Q)$ and $\mu_{true} = \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$.

Based on the quantities defined above, we immediately obtain closed form expressions for $\text{bias}(\hat{\mu}_S) = \omega^t(I_{p-1} - G_S)\delta$ (where we define I_{p-1} as the $(p-1) \times (p-1)$ identity matrix), as well as for $\text{Var}(\hat{\mu}_S)$. Adding squared bias and variance, the limiting expression for $\text{MSE}(\hat{\mu}_S)$ is

$$\text{MSE}(\hat{\mu}_S) = \left(\frac{\partial \mu}{\partial \theta}\right)^t J^{-1}_{00} \frac{\partial \mu}{\partial \theta} + \omega^t Q^0_S \omega + \omega^t (I_{p-1} - G_S)\delta\delta^t (I_{p-1} - G_S)^t \omega.$$

An asymptotically unbiased estimator is obtained by plugging in the sample version of the unknown quantities, and evaluating $\hat{J}$, $\frac{\partial \hat{\mu}}{\partial \theta}$ and $\frac{\partial \hat{\mu}}{\partial \gamma}$ using estimates in the wide model, with $\hat{\delta} = \sqrt{n}\hat{\gamma}$. This leads to the estimated MSE,

$$\widehat{\text{MSE}}(\hat{\mu}_S) = \left(\frac{\partial \mu}{\partial \theta}\right)^t \hat{J}^{-1}_{00} \frac{\partial \mu}{\partial \theta} + 2\hat{\omega}^t \hat{Q}^0_S \hat{\omega} + \hat{\omega}^t (I_{p-1} - \hat{G}_S)\hat{\delta}\hat{\delta}^t (I_{p-1} - \hat{G}_S)^t \hat{\omega} - \hat{\omega}^t \hat{Q} \hat{\omega}. \tag{3}$$

The two terms in the middle of (3) form the estimated focused information criterion, while the first and last term, for a given node, do not depend on the model S and thus do not depend on which parents are chosen.

For a nodewise regression setting, the empirical versions Q_n and J_n of the matrices Q and J are calculated as $Q_n = \sigma^2(n^{-1}Z^T(I - U(U^TU)^{-1}U^T)Z)^{-1}$,

$$J_n = n^{-1}\sum_{i=1}^{n}\frac{1}{\sigma^2}\begin{pmatrix}2 & 0\\ 0 & \Sigma_n\end{pmatrix}; \Sigma_n = n^{-1}\begin{pmatrix}U^TU & U^TZ\\ Z^TU & Z^TZ\end{pmatrix}.$$

where U denotes the matrix of 'protected parents' (by which we mean variables that are considered *a priori* as parents and which are never subject to selection, alongside the model's intercept) and Z denotes the matrix of 'unprotected parents' (by which we mean variables that are considered as potential parents and which are subject to selection (or exclusion), based on the improvement (or lack of improvement) in MSE scores).

A subsequent derivation reveals $\omega = Z^TU(U^TU)^{-1}u - z$, where u and z are the prespecified covariate values corresponding to U and Z. In the example in Section 2, see (1), the covariate values refer to the grades for the top/bottom student on the remaining four variables: MEC, VEC, ANL and STA.

Since the limiting and estimated MSE are defined per node, we next define the FIC for the overall estimated graph, as the nodewise summation of MSEs, where each node $l \in \mathcal{V}$ has a particular model S_l based on which we have estimated $\hat{\mu}_{l;S_l}$,

$$\text{FIC}(\mathcal{G}(\mathcal{E}_\mathcal{S}, \mathcal{V})) = \sum_{l=1}^{p}\widehat{\text{MSE}}(\hat{\mu}_{l;S_l}), \tag{4}$$

where $\mathcal{S} = \{S_1, \ldots, S_p | S_1 \subseteq \{\mathcal{V} \setminus 1\}; \ldots; S_p \subseteq \{\mathcal{V} \setminus p\}\}$. The objective is to minimize (4) over $\mathcal{S}$ such that, depending on the context, $\mathcal{G}$ is a DAG, MN or AG.

4.2 Computational-related aspects

Similar to other scoring-based procedures for graphical models, a search for the best model can be performed by complete enumeration of the graphs and by calculating the corresponding score, followed by testing the selected graphs for DAG, MN or AG requirements. Since the number of possible combinations of the models is of exponential order, this exhaustive procedure can only work for small networks, which renders it impractical for large networks. We have implemented in this study the FIC model selection criterion with local optimum search algorithms such as the hill-climbing approach (to have more comparable search procedures with the competitor criteria), but it is very well possible to combine the FIC with different search algorithms.

The hill-climbing procedure starts from an empty graph and adds, in a first step, the best directed single edge in the graph, according to a corresponding score (in our case, the FIC). We continue from the graph obtained in the first step. If the graph contains at least one edge, a decision among three possible modifications is made: either add an extra arrow between a pair of the remaining nodes, delete one of the present edges, or reverse it. After evaluating all such possible changes, the graph with the smallest FIC value is retained in this step 2. The same process continues until the scores cannot be improved anymore, each time starting from the graph retained in the previous step. The algorithm can be seen as a 'greedy' search strategy. In all numerical calculations performed in this paper, when dealing with DAGs we have implemented the hill-climbing approach under the acyclicity constraint; that is moving from graph $\mathcal{G}$ to $\mathcal{G}'$ by edge addition/deletion/inversion is allowed if $\mathcal{G}'$ is still a DAG and if out of all possible other edge additions/deletions/inversions that can take place, the proposed one leads to the smallest estimated MSE of the entire graph. As such, the global acyclicity constraint is satisfied by choosing at each step an edge coming from the set which maintains this property. Because of this constraint, locally at a node a possible sub-optimal decision may be made in terms of MSE, but the introduced edge is chosen such that the deviation from the best local model is as small as possible without violating the DAG constraints.

Since the FIC score depends on the ω vector, which is modified each time a different value is specified for u or z, different FIC scores may be obtained for different covariate positions, thus for different subjects. To provide more insight into why different focuses may give rise to possibly different models being selected, we concentrate on a simple subproblem based on the 'Mathematics marks' standardized dataset. In order to be able to provide a graphical representation, assume that one has at his disposal only three nodes:

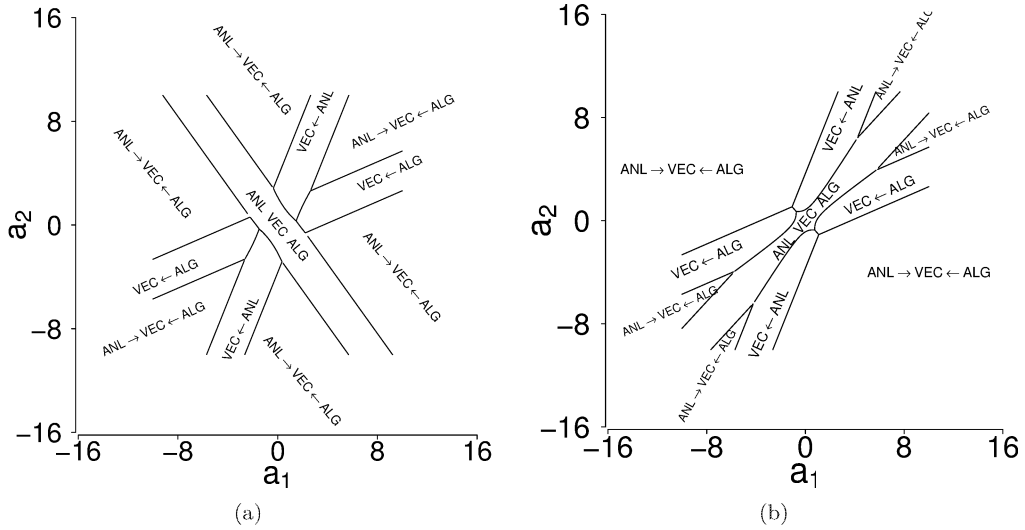


Fig. 3: Mathematics grades. Regions with minimal MSE for four models (see text) with two different focuses. For (a) $\omega = (2.7861, -2.5383)$ was chosen, while for (b) $\omega = (-1.5433, -1.3687)$.

VEC, ALG and ANL, where VEC is a child node and no directed edge exists between ALG and ANL. Model selection is to be performed between four possible models: (i) VEC $\leftarrow$ ALG, (ii) VEC $\leftarrow$ ANL, (iii) ANL $\rightarrow$ VEC $\leftarrow$ ALG and (iv) the empty model where none of the two nodes acts as a parent for VEC. Leaving out terms in the MSE that do not depend on the model used (see Claeskens and Hjort, 2008b, eqn. (6.1)) we obtain the following expressions for the resulting FIC quantities. Model (i) is to be selected when its theoretical FIC value $\mathrm{FIC}_{\mathrm{ALG}} = (\omega^t(I_2 - G_{\mathrm{ALG}})D)^2 + 2\omega^t Q^0_{\{\mathrm{ALG}\}}\omega$ is smaller than the FIC values of all three other models. The $\mathrm{FIC}_{\mathrm{ANL}}$ for model (ii) is defined analogously, replacing ALG by ANL. The FIC for the full model (iii) is equal to $2\omega^t Q\omega$, while the FIC for the empty model (iv) equals $(\omega^t D)^2$. Working out the inequalities that state that one model is better than all other models in terms of FIC value, defines regions in terms of D, that depend on the focus through the value of ω. Since $D \sim N(\delta, Q)$, we replace in the limit experiment D by δ and work with the standardized value $a = Q^{-1/2}\delta$, or equivalently, phrase the inequalities in terms of $Q^{1/2}a$ instead of working with D.

Figure 3 depicts the regions in terms of a where each of the four models attains the lowest MSE. For these calculations we took $\mu = E[\mathrm{VEC}|\mathrm{ALG}, \mathrm{ANL}]$ with in panel (a) the values for ALG and ANL for a student with ω equal to (2.7861,-2.5383) and in panel (b) the corresponding values for a student with ω equal to (-1.5433,-1.3687). We have used the estimated 2×2 matrix Q based on the mathematics grade data only using VEC, ANL and ALG, resulting in the value 1.2735 on the diagonal and -0.9052 off-diagonal. Two important conclusions can be drawn. First, comparing panels (a) and (b), for different focuses (thus different ω vectors) the regions of best performance for a model can be smaller/larger. For a certain value of δ (and thus a), it might be that for the student chosen to draw panel (a) we need to choose the full model, while with that same value of δ for the other student, model (i) VEC $\leftarrow$ ALG, will give the lowest MSE when estimating μ. A second observation from this figure is that the best performing model in terms of MSE for estimation of μ does not need to be the model that has generated the data. In light of this problem, see also Hjort (1994) and Chapter 5 of Claeskens and Hjort (2008b) where several examples are given where a simpler model might give better MSE values than the more complex true model. For example, a regression model might be truly quadratic in one covariate, but when the corresponding coefficient is small in value, the bias from fitting a linear model might not outweigh the added variance by fitting a quadratic model. A similar situation arises in the graphical models, adding a true arrow might reduce the bias, but at the same time might increase the variance such that the resulting MSE of this true model is larger than that of the simpler misspecified model. Figure 3 depicts the regions in the parameter space

for the choice between four models, where each time the MSE of that model is smaller than that of the competing models.

Judging by the relative size of the drawn regions we conclude that for a large set of δ values the model ANL $\rightarrow$ VEC $\leftarrow$ ALG would attain smaller MSE for the studied focus in the situation of (b), whereas for the student in situation (a) the set of values supporting such a model is smaller.

For an application of the FIC in a different multivariate setting, with another choice of focus, see Section 6.2.

Next, we phrase a theoretical result that finds the conditions under which the MSE of the focus estimator for a graph decreases when one arrow is added. The graph with the added arrow ($\mathcal{H}$) is thus the better graph. Note that adding an edge does not always have to lead to a smaller MSE value since the estimation of the additional parameters might cause the variance to inflate, possibly making the MSE of the larger graph larger than that of the graph without this extra edge.

Proposition 1 *Suppose DAG $\mathcal{H}$ extends $\mathcal{G}$ by adding the edge $j \leftarrow i$. If for node j, $|\omega^t \delta| > \sqrt{\omega^t Q^{-1} \omega}$ then $\sum_{l=1}^{p} MSE_{\mathcal{G}}(\hat{\mu}_{l;S_l}) > \sum_{l=1}^{p} MSE_{\mathcal{H}}(\hat{\mu}_{l;S_l})$.*

Proof The proof follows from Theorem 5.3 of Claeskens and Hjort (2008b, p. 127) since the inequality is based on a decomposable criterion, and since the two graphs are identical with respect to all other relations (that is, the nodewise MSEs will be identical for the identical relations). Thus the stated inequality will hold if for node j, $\text{MSE}_{\mathcal{G}}(\hat{\mu}_{j;pa_{\mathcal{G}}(j)}) > \text{MSE}_{\mathcal{H}}(\hat{\mu}_{j;\{i,pa_{\mathcal{G}}(j)\}})$, and this is true if the condition in the lemma holds; otherwise if the condition does not hold then adding the arrow will not necessarily decrease the mean squared error.

Lemma 1, similar to Theorem 3.7 in Williamson (2005, p. 24), justifies theoretically for BNs the circumstances under which adding an arrow does not increase the Kullback-Leibler distance to the target distribution.

Lemma 1 *Let the DAGs $\mathcal{G}$ and $\mathcal{H}$ with corresponding pdfs $f_{\mathcal{G}}$ and $f_{\mathcal{H}}$ be such that $\mathcal{H}$ differs from $\mathcal{G}$ by only one extra directed arrow between nodes i and j (say, $j \leftarrow i$). If for a finite set of graphs there exists a unique target distribution $f^\star$ in the same model class, that minimizes $\sum_{l=1}^{p} MSE(\hat{\mu}_{l;S_l})$, then $f_{\mathcal{H}}$ will be closer to $f^\star$ in terms of Kullback-Leiber distance if there exist a set of values x_j with positive Lebesgue measure such that $f(x_j|x_i, pa_{\mathcal{G}}(x_j)) \neq f(x_j|pa_{\mathcal{G}}(x_j))$ and X_i does not equal a point mass at 0. Otherwise, $f_{\mathcal{H}}$ will not be further away from $f^\star$ than $f_{\mathcal{G}}$.*

Proof Let $d(f^\star, f_{\mathcal{H}})$ be the Kullback-Leiber (KL) distance between the target and $f_{\mathcal{H}}$. Since $\log(a) \leq a - 1 \ \forall a \in \mathbb{R}$, the difference between the two distances is bounded above by

$$d(f^\star, f_{\mathcal{H}}) - d(f^\star, f_{\mathcal{G}}) = \int f^\star(x) \log \frac{f_{\mathcal{G}}(x)}{f_{\mathcal{H}}(x)} dx \leq \int f^\star(x) \frac{f_{\mathcal{G}}(x)}{f_{\mathcal{H}}(x)} dx - 1.$$

Since the networks are identical with respect to all but nodes i and j, $\frac{f_{\mathcal{G}}(x)}{f_{\mathcal{H}}(x)} = \frac{f(x_j|pa_{\mathcal{G}}(x_j))}{f(x_j|x_i, pa_{\mathcal{G}}(x_j))}$ with $x = (x_1, \ldots, x_p)$, thus

$$\int f^\star(x) \frac{f_{\mathcal{G}}(x)}{f_{\mathcal{H}}(x)} dx = \int f(x_i) f(pa_{\mathcal{G}}(x_j)|x_i) f(x_j|pa_{\mathcal{G}}(x_j)) dx = 1,$$

since the last expression leads always to a valid BN over a multivariate distribution which integrates to 1.

To obtain the strict inequality $\log\left(f_{\mathcal{G}}(x)/f_{\mathcal{H}}(x)\right) < f_{\mathcal{G}}(x)/f_{\mathcal{H}}(x) - 1$, use that $\log(a) < a - 1 \Leftrightarrow a \neq 1$ and the factorization of the joint density according to both $\mathcal{G}$ and $\mathcal{H}$. Then $\frac{f(x_j|pa_{\mathcal{G}}(x_j))}{f(x_j|x_i, pa_{\mathcal{G}}(x_j))} \neq 1$ if and only if the densities are different on a set with positive Lebesgue measure.

For normal pdfs $f_{\mathcal{G}}$ and $f_{\mathcal{H}}$ the strict inequality in the proof holds if the true regression coefficient $\beta_{ji} \neq 0$ (see equation 2) since this implies using two normal distributions centered at different locations when X_i is not a point mass at zero. Because of this, and if the conditions in Proposition 1 hold then the inclusion of X_i in the parental set of X_j decreases the corresponding $\text{MSE}_{\mathcal{H}}(\hat{\mu}_{j;\{i,pa_{\mathcal{G}}(j)\}})$.

A pseudo-code algorithm for the implementation of the procedure described in this article can be found in Algorithm 1. In its simplest form, once the focus under study has been chosen and all necessary quantities from equation 3 have been computed, the algorithm starts from an empty graph and updates its structure according to a hill-climbing procedure if the updated graph improves the current estimated score.

Algorithm 1 FIC based search method for BNs

$\hat{\mathcal{G}} \leftarrow$ `empty graph`
$FIC_{\hat{\mathcal{G}}} \leftarrow$ `compute FIC score based on` $\hat{\mathcal{G}}$;
$Flag \leftarrow False$;
while $Flag == False$ **do**
  `compute` Add `based on` $\hat{\mathcal{G}}$;
  `compute` $Delete$ `based on` $\hat{\mathcal{G}}$;
  `compute` $Invert$ `based on` $\hat{\mathcal{G}}$;
  $Allmoves \leftarrow$ `append` $Add, Delete, Invert$;
  $Length \leftarrow$ `the length of` $Allmoves$;
  for $Adjacency = 1 \rightarrow Length$ **do**
    $FIC[Adjacency] \leftarrow$ `compute FIC score based on` $Adjacency$;
  end for
  if $minimum(FIC) < FIC_{\hat{\mathcal{G}}}$ **then**
    $position \leftarrow$ `position of` $minimum(FIC)$;
    $\hat{\mathcal{G}} \leftarrow Adjacency[position]$;
    $FIC_{\hat{\mathcal{G}}} = minimum(FIC)$;
  else
    $Flag \leftarrow True$;
  end if
end while

5 Numerical Results: Simulated Datasets

After a short description of the competitive methods and settings that were created for comparison, we present the results obtained on simulated data in the BN and MN case. Afterwards, all methods are applied on benchmark datasets.

5.1 Competitive Methods for estimating BNs

For comparison purposes, some competing algorithms have been tested, and in the case of BNs we have applied the hill climbing (HC) algorithm in conjunction with BGe score (Heckerman and Geiger, 1995), the PC algorithm (Spirtes et al, 2000) using partial correlation tests for independence testing and as well the 'conditional log-likelihood' maximization proposed in Grossman and Domingos (2004).

For HC we have used the implementation offered in the 'bnlearn' (Scutari, 2010) library for the R software with default options, which meant that in the BGe case for the parameters a Gaussian-inverse Wishart prior was used and a priori each DAG configuration was considered equally likely. The initial parameters are found conditioned upon an 'imaginary sample size' which we took to be 10. Experimentally, Gammelgaard Bøttcher (2004) found this approach to provide close results to the original approach in Heckerman and Geiger (1995) while being simpler to specify.

The PC algorithm starts from a fully connected undirected graph and tests the existence of a directed arrow by testing for independence. Next, orientations are found according to specific rules. This results in a partial DAG (PDAG) as for some of the undirected edges the directionality cannot be uniquely determined. The implementation offered in the 'pcalg' (Kalisch et al, 2012) library for the R software, has been used in this study.

In addition, the algorithm of Dor and Tarsi (1992) which takes the PDAG as input and returns a DAG if the partial DAG admits such an extension (otherwise the algorithm returns the PDAG) has been used to decide on a possible orientation in case of unoriented edges for the PC and BGe estimated DAGs that we present in the paper.

In the context of classification, Grossman and Domingos (2004) proposed the use of maximization of the conditional likelihood instead of maximizing the usual likelihood of the data. They conjectured that since the log-likelihood can be separated into a conditional and a remainder part (which generally is much larger and can thus swamp the contribution of the first term), optimizing such a conditional likelihood would be a better choice. Their experimental results seem to suggest that for the classification problem, such an approach might have merit.

5.2 Competitive Methods for MNs

For the MN setting we have compared with SIN (Drton and Perlman, 2004), graphical lasso (Friedman et al, 2008) and an extended version of the Chow and Liu (1968) procedure.

Drton and Perlman (2004) introduced the SIN methodology for MNs, that was later extended to DAGs and chain graphs in Drton and Perlman (2008), which is based on simultaneous testing whether all partial correlations ρ_{ij} between pairs of variables (X_i, X_j) conditioned on all remaining variables $X_{\mathcal{V}\setminus\{i,j\}}$ are zero. The size of the obtained p-values determines a separation in three groups: *S*ignificant (edges that should be included in the estimated $\mathcal{G}$), *I*ndeterminate (edges for which inclusion might be a decision if a less conservative level were to be chosen) and *N*on-significant (edges that should not be included). The 'SIN' library for the R software, has been used in this study.

Abreu et al (2010) implemented an extended algorithm of Chow and Liu (1968) to estimate the structure of the underlying graph. It proceeds in two steps, where one first estimates a tree/forest structure based on a particular scoring criterion (in these simulations BIC was used) as only edges that preserve the tree/forest structure are added, and then in a forward search to the final model from the first stage, edges are added (based on BIC scores) such that the final obtained model is decomposable and preserves the isolation structure of the model from the first stage. The 'gRapHD' library for the R software, has been used.

In the context of high-dimensional data ($n < p$), Meinshausen and Bühlmann (2006) introduced a procedure based on a series of nodewise lasso regressions to obtain an estimate of a Markov network. Later, Friedman et al (2008) and Witten et al (2011) extended the lasso procedure to multivariate density estimation, by using ideas of penalization directly on the Σ^{-1} matrix, and labeled the resulting procedure as graphical lasso (GLasso), concluding as in Yuan and Lin (2007) that the nodewise procedure is just an approximation to the GLasso. GLasso maximizes the penalized log-likelihood of the data, which is up to a constant,

$$\log\det\Sigma^{-1} - \text{trace}(W\Sigma^{-1}) - \lambda||\Sigma^{-1}||_1,$$

over positive definite matrices Σ^{-1}, where W is the sample covariance matrix and $||\Sigma^{-1}||_1$ is the l_1 norm i.e., the sum of absolute values of the entries in the Σ^{-1} matrix. Varying λ, the amount of penalization on the l_1 norm, causes some elements of the matrix to be set to 0. A Kullback-Leibler loss, $\text{KL} = -\log(|\hat{\Sigma}^{-1}_{\text{train}}|) + \text{tr}(\hat{\Sigma}^{-1}_{\text{train}}\hat{\Sigma}_{\text{test}})$, has been used to select the regularization parameter, where $\hat{\Sigma}^{-1}_{\text{train}}$ is the concentration matrix estimated on the training sample and $\hat{\Sigma}_{\text{test}}$ is the covariance matrix fitted using the test set. Three-fold cross validation has been applied. The 'rotation information criterion' (RIC) implemented in Zhao et al (2012) has also been used for model selection purposes. The 'glasso' and 'huge' libraries for the R software, have been used in this study, with default options and parameter values.

Next we mention other l_1 approaches that were not considered in the simulation study. Schmidt et al (2007), Banerjee et al (2008), Li and Toh (2010) and Krishnamurthy et al (2012) among others, also proposed different variants of using either nodewise l_1 or multivariate l_1 penalized methods to estimate graphs, mostly undirected ones, or for the explicit purpose of estimating Markov blankets of nodes (or removing spurious edges from it). All of these procedures share more or less the same principle of estimating sparse models either as a goal in itself or as an intermediate device, through the use of the l_1 enforced sparsity.

5.3 Simulation Study Setup

Different scenarios have been constructed using sample sizes equal to either 25, 100 or 300 and networks having either 10 or 20 nodes ($X_1, \ldots, X_p$, where $p = 10$ or 20). The structure of the network is completely random and the number of neighbor nodes of any X_j ($j > 1$) is drawn from a binomial distribution with the number of trials equal to $j - 1$, and with the probability of success, ν equal to the probability of connecting two nodes, where $\nu = 0.5$ or 0.7. The neighbors of X_j are drawn with replacement from the set $\{X_1, \ldots, X_{j-1}\}$. Edge weights are drawn from a uniform(0.1, 1) distribution. For each generated network, random errors are generated either from a multivariate normal or a multivariate t-distribution with 10 degrees of freedom and with one of six different covariance matrices. For each setting we took 500 simulation runs and for each dataset both a DAG and a MN have been selected.

Denote by $c_{i,j}$ a matrix entry from row i and column j with $i, j = 1, \ldots, p$. The following models have been used for constructing the covariance matrix:

- 'Banded' : with $c_{i,i} = 1$ and $c_{i,i-1} = c_{i-1,i} = 0.5$
- 'NoCorr' : with $c_{i,i} = 1$
- 'Full' : with $c_{i,i} = 1$ and $c_{i,j} = 0.2\ \forall i \neq j$
- 'Decay' : with $c_{i,j} = 0.9^{|i-j|}$
- 'Cross' : with $c_{i,i} = 1.5$ and $c_{i,p+1-i} = 0.5$
- 'Star' : with $c_{i,i} = 1$ and $c_{1,j} = c_{i,1} = 0.2$

All other matrix entries are set to 0.

Seven focus points $\mu = E[Y|x]$, i.e. the conditional expectation of a node given its parents, have been used in the analysis:

Focus 1 : evaluate μ at the x value of one of the 'in-sample' data points;

Focus 2 : evaluate μ at an 'out-of-sample' data point that comes from the same distribution as used to generate the dataset;

Focus 3 : evaluate μ at an 'out-of-sample' data point that comes from the same distribution as used to generate the dataset, this value stays fixed regardless of the simulation run;

Focus 4 : evaluate μ at an 'in-sample' data point that has the lowest data depth score (thus corresponding to a value on the 'boundary');

Focus 5 : evaluate μ at an 'in-sample' data point that has the largest data depth score (thus corresponding to a median value, in the center);

Focus 6 : evaluate μ at an 'out-of-sample' data point that comes from a t(4) distribution;

Focus 7 : evaluate μ at an 'out-of-sample' data point that comes from a t(4) distribution, this value stays fixed regardless of the simulation run;

Since the datasets have been randomly generated, one might not have as clear focuses such as the scores for a 'top/bottom' ranked student as in Section 2. For a proper use of focused search methods, the focus should be determined before the rest of the analysis (to avoid 'data snooping'). The focus should be clear from the research question. If interest is in prediction, the conditional expectations are an obvious choice. One may think of the chosen focus points as containing the characteristics of a (new) person for which a model to estimate μ is to be selected. By this wide choice of focus points we have challenged the performance of FIC under a multitude of 'diverse interests'. The random generation of the focuses, as opposed to a specification of a precisely predefined focus has in this simulation study the purpose of offering a fair comparison by avoiding the pre-selection or presentation of 'good' settings only.

The estimated models are evaluated in terms of the empirical mean squared error MSE($\mathcal{G}$), the empirical mean squared prediction error MSPE($\mathcal{G}$), the sparsity index (1−the number of estimated edges divided by the total number of possible edges), the F_1 score (Jardine and van Rijsbergen, 1971), the Hamming distance (HD, Tsamardinos et al, 2006) and the precision value for a given fixed value of recall. The recall (or true positive rate, TPR) and precision values are defined as Recall = TP/(TP + FN) and Precision = TP/(TP + FP); where TP denotes the number of correctly found edges by the estimated graph, FP denotes the number of incorrectly found edges, and FN denotes the number of true edges that are not present in the estimated graph. Based on the above rates the F_1 score is defined as: F_1 = 2·Precision·Recall/(Precision+Recall) and by construction it is bounded between 0 and 1, with higher values denoting better graph reconstruction performance.

To construct the precision-recall plots we took the following approach. Out of all estimated graphs from all simulations, we searched for the values for which the recall is the same (up to two decimal places) for FIC and competitors. At this value for recall we then look at the vector of estimated precision values and take the corresponding average precision. Thus, for a given fixed recall rate of, say 0.50, plotted on the horizontal axis, some estimated graphs which attain this recall rate, might have higher or lower precision rates of which we take the average and then plot on the vertical axis.

We define the empirical MSE and MSPE as $\overline{\text{MSE}}(\mathcal{G}) = \sum_{l=1}^{p} \overline{\text{MSE}_l}$ and $\overline{\text{MSPE}}(\mathcal{G}) = \sum_{l=1}^{p} \overline{\text{MSPE}_l}$ where

$$\overline{\text{MSE}_l} = \frac{1}{500} \sum_{k=1}^{500} \Big(\sum_{i \in pa(l)} \beta_{li}^{true} x_{0i}^{(k)} - \sum_{i \in pa(l)} \hat{\beta}_{li} X_i^{(k)} \Big)^2;$$

$$\overline{\text{MSPE}}_l = \frac{1}{500}\sum_{k=1}^{500}\Big(x_{0l}^{(k)} - \sum_{i\in pa(l)}\hat{\beta}_{li}X_i\Big)^2.$$

with the superscript (k) indicating the simulation run and $x_0^{(k)}$ the focus evaluation point. For focuses 3 and 7, $x_0^{(k)}$ is the same for all k.

We need to stress that the competitor methods are performing model selection with the explicit intent to produce graphical structures that come close to the true underlying graph, whereas the FIC method used here optimizes the estimated MSE of the conditional expectation of a node allowing for different edge configurations. The purpose is to evaluate if for such focus functions an FIC selected graph performs better than traditional methods and in what situations. To avoid confusion we stress also that the FIC graphs are all selected based on the estimated MSE of the conditional expectation (3 is used only for estimating a graph) whereas the performance evaluation is based on the actual empirical MSE.

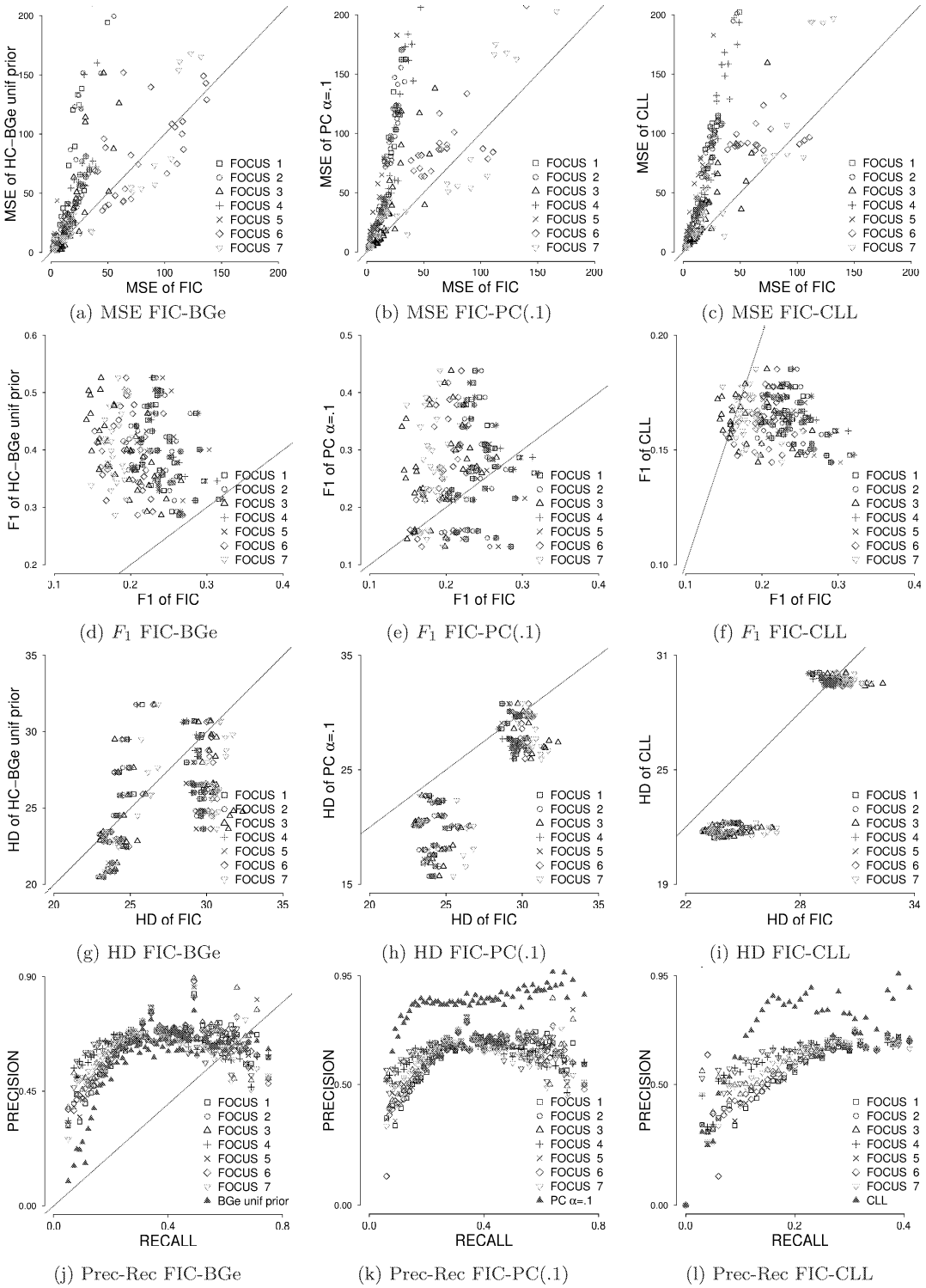
5.4 Simulation Study Results: BNs

In Figure 4 we have plotted the values of the empirical MSE, F_1 score, Hamming distance (HD), and the precision rates for fixed recall rates for FIC versus, the corresponding values for hill-climbing with BGe, the PC algorithm with $\alpha = 0.1$ and the conditional log-likelihood (CLL). For MSE, F_1 and HD each symbol in the plot represents an average over 500 simulation runs. Different symbols correspond to the seven different focuses. Per focus we have 36 settings (3 sample sizes, 6 covariance matrices and 2 values for ν) corresponding to 36 points using the same symbol in the figure. Values above the diagonal indicate smaller values for FIC than for the technique on the vertical axis. Since the results and conclusions are similar for the situations of random errors from a multivariate normal and a multivariate t(10) distribution, we present here only the results for the t(10) distribution. There is also a large degree of similarity between the results obtained from AIC and BIC criteria and BGe and due to space limitations out of all three criteria only the BGe estimated graph summaries are presented in the paper. In the figures, networks of 10 nodes are considered.

With respect to the empirical MSE values, for most of the simulation settings, the values appear above the diagonal in the plot, indicating a favorable position for the FIC. Comparing FIC to the results of the PC algorithm and to that of the CLL approach, in the majority of cases the FIC is able to produce much smaller empirical MSEs. In some cases, the obtained empirical MSEs can be as much as 2–3 times smaller. When compared to the HC based on BGe, the obtained improvements are not as dramatic as previously, but they can still be quite substantial. For only a few settings, for the 3rd, 6th and 7th focus sometimes the empirical MSEs were slightly larger for FIC than for the other methods, especially when compared to BGe. Considering the empirical MSPE instead of the MSE, leads to the same conclusions. These results are not shown here.

In terms of sparsity of the skeleton, the proposed FIC method produces sparser graphs when compared to HC-BGe or CLL, but when compared to the PC the results are inconclusive (not shown here). The FIC estimated networks tended to have generally higher HD scores and lower F_1 scores when compared to the estimated networks from competitor techniques. In several settings the CLL or PC graphs were better in terms of precision rate, while for some settings the FIC model provided precision rates comparable to those of BGe. Here it becomes clear that the objective function in the FIC determines the selected model. By minimizing the MSE, the closeness to the underlying graph is sacrificed for better prediction.

One additional remark with respect to the properties of the skeleton might be appropriate. An alternative use of the FIC is to not use it for a single focus, but to select values based on FIC averages computed for several focus points, resulting in an 'average' FIC model. By an average model we mean that we consider a number (say 5, or 10) different focus values, obtain the FIC score for each focus value, take the average of the FIC values over the focus points for each of the considered models in the search procedure and finally select that model for which the obtained average FIC score is the lowest. Such a model is found to be less sparse (not shown here), and to have both increased recall and increased FPR, but with no important gains in precision. The advantage of an averaged model is that the focus has a somewhat broader interpretation, because, instead of evaluating μ in a single point x, a set of values (e.g. based on quantiles) can be considered, making the selected model preferred for a wider range of x values.



(a) MSE FIC-BGe (b) MSE FIC-PC(.1) (c) MSE FIC-CLL

(d) F_1 FIC-BGe (e) F_1 FIC-PC(.1) (f) F_1 FIC-CLL

(g) HD FIC-BGe (h) HD FIC-PC(.1) (i) HD FIC-CLL

(j) Prec-Rec FIC-BGe (k) Prec-Rec FIC-PC(.1) (l) Prec-Rec FIC-CLL

Fig. 4: Empirical MSE (a-c), F_1 index (d-f), Hamming distance (g-i) and precision-recall plots (j-l) of FIC plotted against the performance of competitor methods for 7 focus values and 36 simulation settings, averaged over 500 simulation runs for networks with 10 nodes. Random errors are generated from a multivariate t(10) distribution.

Table 1: Empirical root-MSE of estimated graphs for three focus points, averaged over 500 simulation runs. Errors are generated from a multivariate normal distribution with different covariance matrices, different sample sizes and values for ν as indicated in the first column. All networks contain 20 nodes.

	Focus 3				Focus 6				Focus 7			
		PC				PC				PC		
Cov(n,ν)	FIC	(.1)	CLL	BGe	FIC	(.1)	CLL	BGe	FIC	(.1)	CLL	BGe
Banded(100,.5)	61	70	70	**52**	120	131	118	**79**	28	29	30	**21**
Banded(100,.7)	671	943	804	**645**	384	478	518	**378**	80	80	87	**71**
Banded(300,.5)	137	**121**	137	138	109	107	123	**96**	13	12	16	**12**
Banded(300,.7)	**204**	262	255	258	**398**	537	522	410	**64**	77	74	77
Banded(25,.5)	36	32	35	**26**	134	128	127	**86**	26	24	28	**21**
Banded(25,.7)	**15**	19	37	27	436	561	594	**354**	407	634	477	**321**
Decay(100,.5)	**77**	127	121	102	**89**	130	125	94	**22**	34	31	23
Decay(100,.7)	**952**	1811	1334	1157	**355**	700	520	496	**57**	89	90	73
Decay(300,.5)	**358**	501	543	535	**87**	97	125	107	**10**	13	29	30
Decay(300,.7)	**642**	1075	1028	1219	**322**	409	497	435	**55**	79	98	117
Decay(25,.5)	64	66	68	**56**	**116**	140	131	120	24	25	35	**23**
Decay(25,.7)	95	124	111	**86**	449	566	636	**418**	387	578	475	**358**
Full(100,.5)	**33**	53	55	38	71	122	131	**68**	22	28	30	**18**
Full(100,.7)	**428**	752	646	434	318	445	496	**309**	**52**	81	87	63
Full(300,.5)	**213**	334	358	302	66	113	121	**63**	**9**	12	17	12
Full(300,.7)	**351**	715	641	588	**290**	575	485	390	**42**	79	77	67
Full(25,.5)	31	29	33	**23**	105	147	127	**83**	23	24	29	**21**
Full(25,.7)	**15**	20	35	25	430	555	526	**376**	404	558	459	**294**
NoCor(100,.5)	**31**	48	50	32	73	123	130	**65**	23	28	28	**16**
NoCor(100,.7)	426	698	601	**419**	334	566	499	**332**	**51**	80	83	64
NoCor(300,.5)	**48**	70	80	61	**66**	105	121	71	**9**	12	16	9
NoCor(300,.7)	**78**	152	146	131	**297**	417	482	411	**42**	79	73	65
NoCor(25,.5)	27	25	28	**19**	104	148	128	**84**	23	24	25	**19**
NoCor(25,.7)	**12**	16	31	24	429	557	531	**327**	413	558	466	**298**
Star(100,.5)	**39**	57	57	39	**75**	116	132	75	25	29	29	**18**
Star(100,.7)	551	773	646	**489**	**362**	566	490	368	**54**	81	82	67
Star(300,.5)	**141**	207	234	197	**72**	93	123	79	**9**	12	20	13
Star(300,.7)	**228**	398	361	366	**333**	523	490	414	**43**	77	79	76
Star(25,.5)	29	27	30	**21**	106	151	130	**92**	23	24	24	**20**
Star(25,.7)	**11**	13	26	23	400	520	505	**339**	464	561	478	**299**

Table 1 contains the empirical root-MSE values for networks with 20 nodes, and errors generated from a multivariate normal distribution, and this for three of the focuses, namely focus 3, 6 and 7. For this larger network, we come to similar conclusions as for the smaller network, summarized in Figure 4. In many settings, the graph selected by FIC has a lower empirical root-MSE for the focus estimator than the graphs obtained from the other methods.

In order to provide an overall summary of the numerous simulation settings, we have performed a small 'meta-analysis' by employing an analysis of variance. Specifically, as response values we use the difference on the log scale between the average empirical root mean squared errors of the estimator in the FIC selected model and that of the competitors. The average has been taken over 500 simulation runs. For the BN case, we compare with hill-climbing using BGe and with the PC algorithm for focus values 3, 6 and 7. Because we have 120 simulation settings, this results in a 'sample size' for the analysis of variance equal to 120 (where we have used 5 covariance structures, 3 sample sizes, 2 values for ν, networks with 10 or 20 nodes and errors from a t(10) or normal distribution). For all categorical variables in the analysis of variance, the category that is not mentioned (e.g. $N = 25$ distribution) is the reference category. Table 3 presents the 95% confidence intervals of the estimated coefficients in the analysis of variance model. Due to the differences that have been calculated, positive estimated values indicate a preference for the FIC.

When compared to a banded covariance structure, any other model increases the difference between the PC estimated graphs and the FIC based ones on the log empirical root-MSE scale, showing a clear advantage for the FIC for all three focus points. In the HC-BGe versus FIC comparison, only for the seventh focus there is an advantage of FIC, for all the rest there is not enough evidence to support the claim. Switching from a multivariate normal to a multivariate t(10) distribution, does not deteriorate the performance of FIC much more that the performance of competitor techniques.

Moving from a sparser network to a denser one (for the same number of nodes), increasing the number of nodes or the number of samples generally increases the empirical root MSE error for the competitors and this is detrimental to a larger extent for PC than for BGe.

Table 2: Average sparsity index of estimated Bayesian and Markov network graphs for 500 simulation runs for networks with 20 nodes. For the BN case, random errors are generated from a multivariate normal distribution, for the MN case a multivariate t(10) distribution has been used.

	BN					MN					
	FIC			PC		FIC			SIN	GL	Fwd
Cov(n,ν)	F3	F6	F7	(.1)	BGe	F3	F6	F7	(.25)	KL	BIC
Banded(100,.5)	.9	.9	.9	.9	.8	.9	.8	.8	.3	.4	.7
Banded(100,.7)	.9	.9	.9	.9	.8	.8	.8	.8	.3	.4	.8
Banded(300,.5)	.8	.8	.8	.8	.6	.8	.8	.8	.2	.4	.6
Banded(300,.7)	.8	.8	.8	.8	.7	.8	.8	.8	.2	.4	.6
Banded(25,.5)	.9	.9	.9	.9	.9	.9	.9	.9	.9	.4	.8
Banded(25,.7)	.9	.9	.9	.9	.9	.9	.9	.9	.9	.5	.8
Decay(100,.5)	.9	.9	.9	.9	.9	.9	.9	.9	.8	.6	.7
Decay(100,.7)	.9	.9	.9	.9	.9	.9	.9	.9	.8	.7	.8
Decay(300,.5)	.9	.9	.9	.8	.8	.9	.9	.9	.5	.6	.5
Decay(300,.7)	.9	.9	.9	.8	.8	.9	.9	.9	.5	.7	.6
Decay(25,.5)	.9	.9	.9	.9	.9	.9	.9	.9	1.0	.6	.8
Decay(25,.7)	.9	.9	.9	.9	.9	.9	.9	.9	1.0	.7	.8
Full(100,.5)	.9	.9	.9	.9	.8	.9	.9	.9	.8	.4	.8
Full(100,.7)	.8	.9	.9	.9	.9	.8	.9	.9	.8	.4	.8
Full(300,.5)	.9	.9	.9	.8	.7	.9	.9	.9	.6	.4	.7
Full(300,.7)	.9	.9	.9	.8	.7	.8	.9	.8	.5	.4	.7
Full(25,.5)	.9	.9	.9	.9	.9	.9	.9	.9	1.0	.4	.9
Full(25,.7)	.9	.9	.9	.9	.9	.9	.9	.9	1.0	.5	.9
NoCor(100,.5)	.9	.9	.9	.9	.8	.9	.9	.9	.8	.3	.8
NoCor(100,.7)	.8	.9	.9	.9	.8	.8	.9	.9	.7	.4	.8
NoCor(300,.5)	.9	.9	.9	.8	.7	.8	.9	.9	.6	.3	.7
NoCor(300,.7)	.8	.9	.9	.9	.7	.8	.8	.8	.4	.4	.7
NoCor(25,.5)	.9	.9	.9	.9	.9	.9	.9	.9	1.0	.6	.8
NoCor(25,.7)	.9	.9	.9	.9	.9	.9	.9	.9	1.0	.7	.8
Star(100,.5)	.8	.9	.8	.9	.8	.8	.9	.8	.6	.3	.8
Star(100,.7)	.7	.9	.9	.9	.8	.7	.9	.9	.5	.4	.8
Star(300,.5)	.9	.9	.9	.8	.6	.9	.8	.9	.4	.3	.7
Star(300,.7)	.8	.9	.8	.9	.7	.7	.8	.8	.2	.4	.7
Star(25,.5)	.9	.9	.9	.9	.9	.9	.9	.9	1.0	.4	.9
Star(25,.7)	.9	.9	.8	.9	.9	.9	.9	.9	1.0	.4	.9

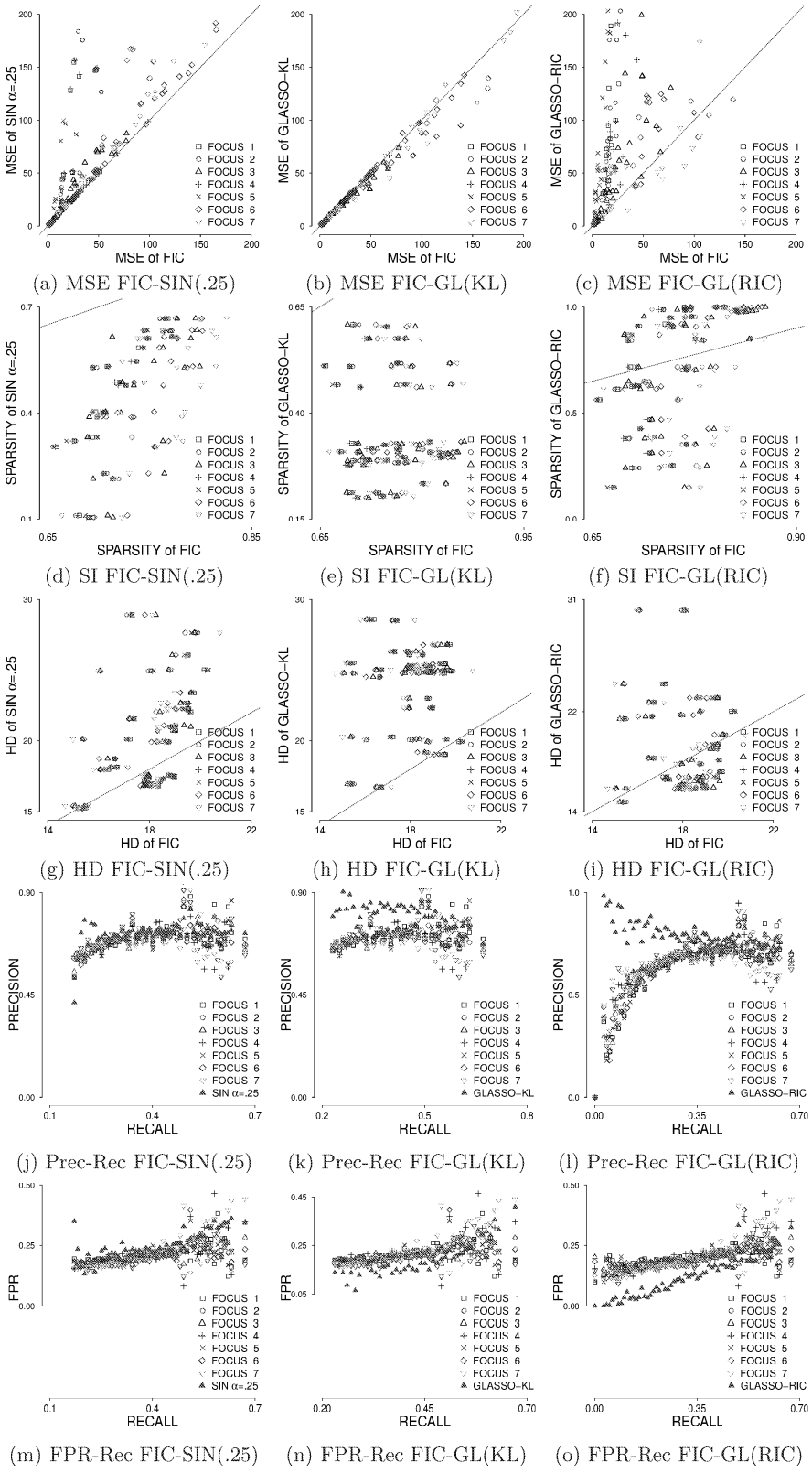
Table 3: Analysis of variance table for a comparison of the different BN simulation settings with respect to the empirical log(root MSE). In parentheses are the 95% confidence intervals for the estimated parameters. Positive estimates indicate a preference for the FIC.

	PC vs FIC			BGe vs FIC		
Focus	3	6	7	3	6	7
β_0	(-.42,-.19)	(-.20,-.04)	(-.39,-.18)	(-.33,.08)	(-.51,.05)	(-.96,-.16)
β_{Decay}	(.25,.46)	(.19,.34)	(.22,.41)	(-.04,.33)	(-.24,.27)	(.50,1.22)
β_{Full}	(.23,.44)	(.29,.44)	(.25,.45)	(-.13,.24)	(-.38,.13)	(.29,1.02)
β_{NoCorr}	(.14,.35)	(.21,.36)	(.25,.44)	(-.18,.20)	(-.39,.13)	(.28,1.01)
β_{Cross}	(.05,.32)	(.14,.32)	(.14,.39)	(-.22,.25)	(-.35,.30)	(.17,1.10)
β_{Star}	(.12,.33)	(.18,.33)	(.19,.38)	(-.22,.16)	(-.40,.12)	(.24,.97)
$\beta_{\text{t(10)}}$	(-.02,.11)	(-.05,.04)	(-.06,.07)	(-.05,.18)	(-.01,.31)	(-.39,.06)
$\beta_{N=100}$	(.05,.22)	(.00,.12)	(.02,.18)	(-.47,-.16)	(-.14,.29)	(-.24,.36)
$\beta_{N=300}$	(.19,.37)	(.01,.13)	(-.10,.06)	(.22,.53)	(.28,.71)	(-.18,.42)
$\beta_{p=20}$	(.10,.23)	(.06,.15)	(.10,.22)	(.02,.26)	(-.08,.25)	(-.35,.12)
$\beta_{\nu=.7}$	(.10,.22)	(.09,.18)	(.17,.29)	(.12,.35)	(-.01,.30)	(-.10,.34)

5.5 Simulation Study Results: MNs

Table 4 and Figure 5 present the values of the empirical MSE, the sparsity index, Hamming distance, the precision rates for fixed recall rates and FPR for fixed TPR for FIC versus the corresponding values for SIN($\alpha = 0.25$), GLasso(KL) and GLasso(RIC). We refer to the description of Figure 4 for an explanation of the plotted values. Also here only the results for errors from a multivariate t(10) distribution are given, the results for the multivariate normal are similar.

In the figure, networks of 10 nodes are considered, and values above the diagonal indicate smaller values for FIC than for the competitor on the vertical axis.



(a) MSE FIC-SIN(.25) (b) MSE FIC-GL(KL) (c) MSE FIC-GL(RIC)

(d) SI FIC-SIN(.25) (e) SI FIC-GL(KL) (f) SI FIC-GL(RIC)

(g) HD FIC-SIN(.25) (h) HD FIC-GL(KL) (i) HD FIC-GL(RIC)

(j) Prec-Rec FIC-SIN(.25) (k) Prec-Rec FIC-GL(KL) (l) Prec-Rec FIC-GL(RIC)

(m) FPR-Rec FIC-SIN(.25) (n) FPR-Rec FIC-GL(KL) (o) FPR-Rec FIC-GL(RIC)

Fig. 5: Empirical MSE (a-c), sparsity index (d-f), Hamming distance (g-i), Precision-Recall plots (j-l) and FPR-Recall plots (m-o) of FIC plotted against the performance of competitor methods for 7 focus values and 36 simulation settings, averaged over 500 simulation runs for networks with 10 nodes. Random errors are generated from a multivariate t(10) distribution.

The impressive MSE improvements seen in the BN case cannot be replicated by the FIC when compared to SIN(.25) or Glasso(KL). In these cases the differences are smaller (see Figure 5 and Table 4), signalizing the fact that for Markov networks there is less room for improvement in terms of MSE (the same holds for MSPE). There is however, a noticeable advantage of FIC over the other two methods, in the sense that for the FIC selected models, even if they have similar empirical MSE performance, less edges appear in the estimated graph and thus a sparser model is used with similar empirical MSE capabilities (see also Table 2).

The GLasso based on RIC seemed to under-perform in terms of empirical MSE, sometimes estimating empty graphs with no connections (for both small and large networks).

Focuses 3 and 7 that were chosen at random and then kept fixed over the simulation runs, seem to be more problematic for selecting a Markov network. This may happen because once a 'bad' point is chosen, then the under-performance will be perpetuated across simulation runs. For Markov networks it depends on the setting whether there is improvement when using FIC, as in some settings the model selected by FIC can perform satisfactory while in others it may under-perform. There is no overall best method. In general, when there is a clear focus, the use of the FIC is advocated. When there is no particular part of the graph that is of interest, another method might better serve the purpose.

In general, in the undirected case the FIC was able to estimate Markov networks for which the Hamming distance was generally lower than that of the competitors, and FIC had smaller precision rates when compared to GL, but roughly similar to those of SIN(α=.25).

We want to stress again that the FIC is constructed to perform well regarding MSE, not for true discoveries. Using an average model did not generally increase the size of the estimated model as in the DAG case, for Markov networks the complexity of the model selected via an averaged FIC remaining roughly the same. Similar conclusions hold for multivariate normal errors.

6 Benchmark datasets

Ten datasets have been used to illustrate the performance of the method for benchmarking purposes. These are: 'Mathematics grades' (Mardia et al, 1979); 'Glucose control' (Cox and Wermuth, 1996); 'Bone mineral content' (Edwards, 2000); 'Fret's heads' dataset (Whittaker, 1990), as well as 'Boston Housing', 'Vowel', 'Ozone' (all from the UCI Machine Learning Repository, `http://archive.ics.uci.edu/ml/`), 'Body Fat' (`http://www.amstat.org/publications/jse/datasets/fat.txt`) and a dataset constructed as in Friedman (1991) using 200 sample cases.

For all cases we show results for the MSPE. For the real datasets we use a leave-one-out (LOO) approach, where for each dataset training was performed using $n-1$ sample cases and evaluation was performed on the remaining case. All variables have been standardized to have mean equal to 0 and variance equal to 1, and models have been estimated on the standardized datasets. The expected value of a child at a fixed position of the parents was used as focus. Each data point has been used as focus point, in an LOO evaluation scheme.

The overall summary of Table 5 is that over all estimated graphs (directed and undirected) the FIC wins more times than all competitors put together when MSPE is concerned. The sparsity index of the selected graphs is given in Table 6 and with respect to this criterion there is no general conclusion immediately apparent, sometimes FIC results in sparser graphs than other methods, sometimes the FIC model is more dense. None of the considered methods is everywhere the sparsest.

6.1 Accommodating a Partially Known Structure and Latent Confounders.

It is not always needed to allow the possibility that any node can be a parent of any other node. For example in a DAG where both age and income of a person are variables of interest, it makes more sense to not allow the occurrence of a relationship such as Income $\rightarrow$ Age (since it might make more sense if the edge was reversed). Any *a priori* knowledge about non-allowable relationships may be incorporated.

The FIC algorithm is easily modified to start from a predefined structure (which for example, might be an undirected skeleton coming from an estimation procedure such as PC). We specify the most complex graph that we are willing to consider, and search through subsets of this graph only. This is obtained by

Table 4: Empirical root-MSE of estimated graphs for three focus points, averaged over 500 simulation runs. Errors are generated from a multivariate t(10) distribution with different covariance matrices, different sample sizes and values for ν as indicated in the first column. All networks contain 20 nodes.

	Focus 3				Focus 6				Focus 7			
Cov(n,ν)	FIC	SIN (.25)	GL KL	Fwd BIC	FIC	SIN (.25)	GL KL	Fwd BIC	FIC	SIN (.25)	GL KL	Fwd BIC
Banded(100,.5)	62	66	54	**49**	113	126	98	**84**	28	29	25	**24**
Banded(100,.7)	702	733	664	**622**	388	410	378	**354**	78	82	60	**51**
Banded(300,.5)	147	180	115	**113**	106	130	92	**89**	13	14	10	**9**
Banded(300,.7)	220	245	190	**189**	393	450	356	**355**	62	68	54	**53**
Banded(25,.5)	18	19	13	**9**	114	133	104	**70**	67	74	53	**40**
Banded(25,.7)	21	26	16	**13**	149	168	114	**89**	91	99	73	**62**
Decay(100,.5)	**81**	89	82	**81**	89	94	84	**78**	**23**	24	23	**23**
Decay(100,.7)	1046	1110	1033	**990**	**340**	355	352	352	57	53	51	**50**
Decay(300,.5)	**389**	441	371	408	92	95	**76**	91	10	10	**8**	9
Decay(300,.7)	714	771	698	**711**	331	351	**313**	335	56	57	**50**	52
Decay(25,.5)	57	93	43	**38**	102	144	79	**68**	62	80	43	**41**
Decay(25,.7)	76	125	58	**54**	127	183	99	**95**	81	113	67	**62**
Full(100,.5)	**36**	38	**36**	37	**73**	80	74	75	**23**	**23**	24	24
Full(100,.7)	481	481	475	**465**	**331**	368	348	336	51	53	50	**47**
Full(300,.5)	**234**	266	248	248	**67**	74	69	71	9	9	**8**	**8**
Full(300,.7)	**395**	410	403	401	**294**	316	313	352	**42**	44	43	44
Full(25,.5)	21	32	15	**13**	94	144	83	**72**	57	81	47	**38**
Full(25,.7)	29	49	22	**19**	113	189	108	**91**	90	119	72	**62**
NoCor(100,.5)	**33**	35	**33**	34	**73**	82	74	77	24	**23**	24	24
NoCor(100,.7)	476	461	448	**439**	339	372	341	**316**	50	53	50	**47**
NoCor(300,.5)	**53**	57	54	54	**69**	72	70	75	9	9	**8**	**8**
NoCor(300,.7)	88	89	85	**84**	**313**	327	313	354	**42**	45	43	43
NoCor(25,.5)	57	93	43	**38**	102	144	79	**68**	62	80	43	**41**
NoCor(25,.7)	76	125	58	**54**	127	183	99	**95**	81	113	67	**62**
Cross(100,.5)	**31**	33	32	**31**	**75**	87	77	78	**24**	**24**	25	**24**
Cross(100,.7)	368	354	347	**340**	353	360	**336**	340	51	53	51	**48**
Cross(300,.5)	**107**	121	114	115	**69**	74	72	78	9	9	**8**	**8**
Cross(300,.7)	202	209	**199**	**199**	**311**	329	314	327	**43**	46	44	44
Cross(25,.5)	10	9	9	**7**	96	142	89	**68**	56	81	51	**39**
Cross(25,.7)	11	26	9	**8**	116	190	117	**87**	92	116	77	**62**
Star(100,.5)	42	42	**38**	**38**	80	81	**75**	79	27	25	**24**	**24**
Star(100,.7)	616	588	484	**473**	378	381	345	**325**	52	57	50	**47**
Star(300,.5)	**155**	181	163	161	73	79	**70**	72	9	9	**8**	**8**
Star(300,.7)	259	275	228	**227**	345	385	**314**	319	44	47	**43**	**43**
Star(25,.5)	10	9	8	**6**	102	144	86	**71**	57	82	48	**39**
Star(25,.7)	10	19	8	**7**	126	188	110	**91**	103	116	74	**62**

Table 5: Ten benchmark datasets. Leave-one-out cross-validation averaged MSPE values for the selected BNs and MNs.

	BN						MN				
Dataset(n, p)	FIC	BIC	AIC	PC (.1)	CLL	BGe	FIC	SIN (.25)	GL KL	GL RIC	Fwd BIC
Friedman1(200,11)	**9.4**	10.3	10.2	10.4	10.3	10.3	9.3	**9.2**	9.4	9.5	9.3
Ozone(203,13)	**6.6**	8.1	8.0	9.0	10.9	8.5	**5.8**	6.5	6.2	6.9	6.4
Bones(139,6)	**4.6**	4.6	4.6	4.7	4.8	4.9	3.8	4.3	3.7	**3.7**	3.8
Glucose(68,5)	4.4	4.3	**4.2**	4.3	4.5	4.4	4.0	4.3	4.0	4.0	**4.0**
Heads(25,4)	2.4	2.3	2.2	2.9	2.9	**2.2**	1.7	1.8	**1.6**	1.6	2.0
Grades(88,5)	3.5	3.2	3.2	3.4	3.8	**2.9**	2.6	2.7	2.7	2.7	**2.6**
Wine(178,13)	**7.1**	8.1	8.0	9.0	11.4	8.9	**6.1**	6.5	6.2	7.8	6.5
Housing(506,14)	**6.8**	7.6	7.5	8.5	12.5	9.1	**5.7**	6.1	6.2	6.1	6.3
Vowel(990,9)	**6.7**	7.2	7.1	7.3	8.5	7.5	**5.7**	5.8	5.9	6.1	5.7
Fat(252,18)	**4.5**	5.8	5.6	9.4	16.0	9.7	**4.1**	6.5	4.3	4.4	4.9

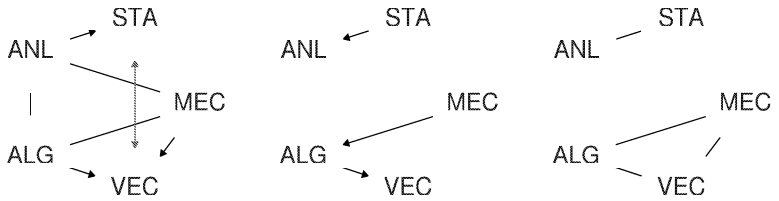
restricting the range of possible combinations of parents that can influence node j, as opposed to evaluating all HC moves. Searching for ancestral graphs that allow the inclusion of latent factors, can also be easily done with the FIC.

We present in Figure 6 three estimated models for one subject from the mathematics grades dataset, see Section 2. We estimate (a) an AG where all edges are allowed between any two nodes, (b) a DAG where we use as starting point the estimated skeleton from the PC(.1) graph, but do not allow the relation ANL $\rightarrow$ ALG $\leftarrow$ STA in the estimated graphs and (c) a MN starting from the same PC(.1) skeleton with

Table 6: Average Sparsity index for the estimated BN and MN graphs for ten benchmark datasets.

	BN						MN				
Dataset	FIC	BIC	AIC	PC (.1)	CLL	BGe	FIC	SIN (.25)	GL KL	GL RIC	Fwd BIC
Friedman1	.74	.91	.69	.83	.80	.80	.83	.84	.74	.85	.84
Ozone	.76	.67	.54	.87	.92	.58	.82	.80	.56	.81	.70
Bones	.60	.60	.54	.67	.80	.59	.71	.80	.28	.26	.67
Glucose	.65	.60	.60	.60	.59	.60	.71	.68	.42	.40	.60
Heads	.42	.35	.29	.61	.61	.30	.56	.71	.00	.00	.41
Grades	.55	.40	.40	.40	.60	.20	.63	.40	.00	.00	.40
Wine	.77	.69	.51	.77	.89	.61	.82	.77	.38	.79	.68
Housing	.77	.54	.37	.75	.95	.56	.80	.64	.49	.45	.69
Vowel	.57	.41	.24	.61	.78	.46	.64	.31	.47	.68	.28
Fat	.82	.67	.55	.83	.89	.71	.86	.83	.37	.39	.75

the same restrictions as the DAG. The estimated PC skeleton used for (b) and (c) imposes edge restrictions as some edges do not appear in the graph (e.g. ANL – MEC – STA). The final estimated DAG suggests that using for example, the relation MEC → ALG generates the smallest MSE estimate for the expected value of ALG, when compared to any other configuration of parents from the allowed set, since any of the others are either too biased (the small models exclude important relations) or produce a large variance (the large models probably include unnecessary edges which inflate the variance of the estimated focus).



(a) (b) (c)

Fig. 6: Mathematics grades. Estimated models for the 13th ranked student: (a) AG with no edge restrictions, (b) DAG and (c) MN using the skeleton from the PC(.1) procedure plus an additional restriction.

6.2 Other Focus Values: Selection From a List of Prespecified Networks

Assuming multivariate normality for the data, we can define focuses on the entire covariance matrix.

Using the 'Mathematics grades' dataset, six focuses have been defined: two generalized standard deviation measures $\mu_1 = (\det \Sigma)^{1/8}$ and $\mu_2 = \sqrt{tr(\Sigma)}$, two partial regression coefficients $\mu_3 = \beta_{\text{STA;MEC}}$ (STA ← MEC) and $\mu_4 = \beta_{\text{ALG;VEC}}$ (i.e. ALG ← VEC), an upper diagonal average measure of correlation $\mu_5 = (1/10)\Sigma_{i<j}\text{corr}(X_i, X_j)$ and μ_6 the correlation between a linear combination of MEC, VEC and ALG and ANL. The goal is to find for these particular focuses of interest, graphs that provide low MSE values when a particular focus is estimated.

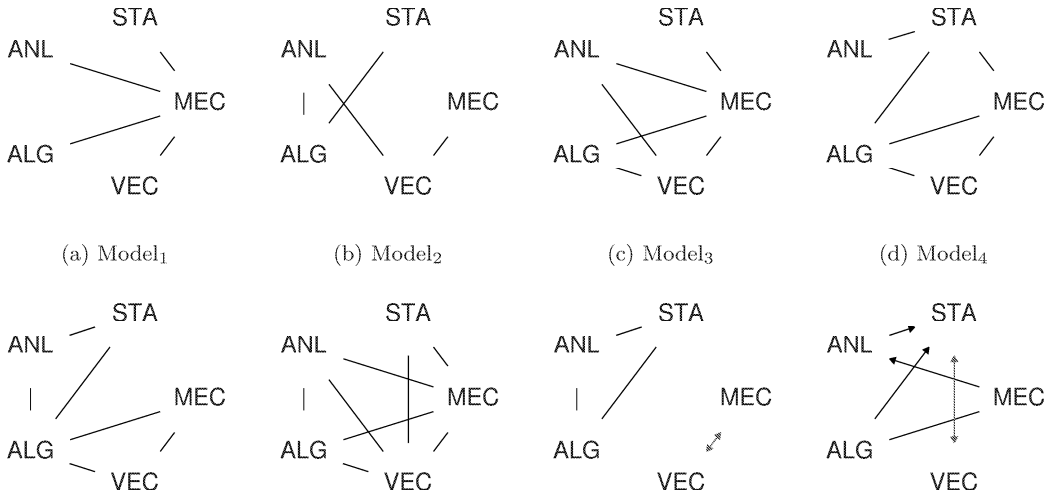
We stress that in this section we use the FIC methodology to select among a few candidate models, the ones that perform best with respect to the estimation of specific focuses. We do not estimate plausible models from the data, but instead build them from theory or by making educated guesses or reasonable assumptions (although an incremental edge addition as in the hill-climbing approach might also be used). Once such a list of candidate models has been constructed, the FIC methodology can aid in the selection of best fitting models (see Claeskens and Hjort, 2008b).

Table 7: Mathematics grades. Focused information criterion values for six focuses and 10 possible multivariate normal models, with different covariance matrices. Models with rank (1) are preferred by the criterion.

Model	FIC(μ_1)		FIC(μ_2)		FIC(μ_3)		FIC(μ_4)		FIC(μ_5)		FIC(μ_6)	
Empty	91.1		292.3		0.2	(1)	38.1		17.4		140.5	
$M1$	43.6		139.8		10.3		38.1		7.6		91.0	
$M2$	4.7	(3)	15.1		0.2	(1)	38.1		0.9	(3)	25.5	
$M3$	26.2		84.1		10.6		10.8	(3)	4.5		54.0	
$M4$	5.0		15.9		5.2		9.5	(2)	0.9		140.5	
$M5$	1.7	(1)	5.3	(1)	0.2	(1)	8.8	(1)	0.3	(1)	14.1	(2)
$M6$	5.4		17.2		8.2		12.9		0.9		20.5	
$M7$	4.8		15.3	(3)	0.2	(1)	38.1		1.0		13.6	(1)
$M8$	15.7		50.4		0.2	(1)	38.1		2.7		90.4	
Full	1.7	(2)	5.5	(2)	6.2		11.8		0.3	(2)	16.2	(3)

Different from the previous application of FIC selection in this paper, we now define focuses that refer to the matrix Σ, and no longer concentrate on the expected value as a focal point. Model selection is performed by checking a sequence of models where particular elements in the concentration matrix are set to 0, while all others are freely estimated.

We use a collection of ten candidate models. Figure 7 contains graphs corresponding to six plausible multivariate 5-dimensional normal densities with mean 0 and different covariance matrices, resulting in different Markov networks, see panels (a)–(f). To these six models, we add an empty graph as well as a fully connected MN (not shown), and two particular AGs, see panels (g) and (h). We let FIC select between these ten models only and present the obtained results in Table 7.



(a) Model$_1$ (b) Model$_2$ (c) Model$_3$ (d) Model$_4$

(e) Model$_5$ (f) Model$_6$ (g) Model$_7$ (h) Model$_8$

Fig. 7: Mathematics grades. Eight candidate models, from which selection of the best model takes place.

For this particular dataset, for all focuses except μ_6, model 5 was selected as the best model with respect to the MSE. For μ_1, μ_2 and μ_5 a fully connected model comes second, and since this model estimates more parameters, it will generally provide larger variances for the focus parameter. For μ_4 model 4 is second best, while for μ_6 the second best model is model 5. This again illustrates that the selected model may depend on the particular focus.

For μ_3, all models that do not set any edge between STA and MEC and have thus, a corresponding 0 entry in the concentration matrix, perform equally well. They all have the same estimated bias and the same estimated variance as the null model, because with respect to μ_3, all 5 models are equivalent. These models seem not to be heavily biased when compared to the full model and so selecting one of these models to estimate μ_3 seems to be the proper way. With respect to μ_4, the first three models and the two ancestral graphs, estimate it at 0 and their scores are as well identical, but these models have large biases that dominate the estimated MSE. Models 4 and 5 are not that biased when compared to the full model, and provide also variances that are close to the variance corresponding to an empty model. Selecting model 5 for the estimation of this particular focus is the best decision, but model 4 follows closely. For the sixth focus, the ancestral graph proposed in Figure 7(g) performs best, among all proposed models.

The important message is that also for focuses defined on the entire covariance matrix it is possible –and worthwhile– to direct the model search towards finding a model with small MSE for the focus of interest. A general second remark is the fact that, as for the traditional BIC and AIC, 'how large is a large difference' between two competing models, is still an issue that requires careful thinking. Selection between models with similar performance might be dictated by other external criteria such as theory, cost or availability of resources. Including all quantities in the estimated MSE expression (hence not leaving out constants not depending on the model) results in FIC values of which the square root is an estimator of the root-mean squared error. This is a quantity with a clear interpretation.

7 Discussion

Global models can often be improved for specific purposes, e.g. in terms of MSE or MSPE. We have defined the focused information criterion for graphical models with the purpose of obtaining a model tailored to a particular objective of the researcher. Depending on the objective or focus, the selected model may be different. The FIC employs an estimate of the mean squared error, explicitly using the trade-off between bias and variance. The advantages of the FIC are that (i) specific model parameters or functions thereof (focus) are allowed to be estimated more accurately than other parts of the model, (ii) ensuing predictions for these focus parameters are generally more accurate, which follows from the low MSE, and (iii) the assumption of the correct model can be relaxed.

We have concentrated at this stage on relatively small networks using Bayesian and Markov networks. Extensions towards larger networks are a topic of current research. The focused scoring criterion is in several instances able to identify graphical models with better empirical MSE and MSPE values than obtained from existing methods, such as the models identified with hill-climbing using AIC and BIC, PC (with additional orientation), SIN and the GLasso (RIC) algorithm. A comparison with hill-climbing using BGe score or with GLasso(KL) depends on the structure of the data and the focus point.

For the BN setting, on several of the considered datasets, the FIC based method is more powerful in terms of leave-one-out crossvalidation error than competitors. On the datasets, in the MN setting, the focused search also performed comparably or better for several of the benchmark datasets. For the simulation settings on which we have tested the algorithm, in the large majority of cases the FIC performs better or comparable in the directed case, while in the undirected case the improvements in empirical MSE were smaller, with the added value of having generally sparser graphs.

Different models can be better in terms of MSE for particular focuses, highlighting once again that the purpose of the model which is reflected in the chosen focus to estimate, is important in the sense that different purposes, and thus different focuses, may lead to different selected models.

Acknowledgements The authors wish to thank the reviewers for their constructive comments. E. Pircalabelu and G. Claeskens acknowledge the support of the Fund of Scientific Research Flanders, KU Leuven grant GOA/12/14 and of the IAP Research Network P7/06 of the Belgian Science Policy. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Hercules Foundation and the Flemish Government - department EWI.

References

Abreu G, Labouriau R, Edwards D (2010) High-dimensional graphical model search with the gRapHD R package. Journal of Statistical Software 37(1):1–18

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov B, Csáki F (eds) Second International Symposium on Information Theory, Akadémiai Kiadó, Budapest, pp 267–281

Ali RA, Richardson T, Spirtes P (2009) Markov equivalence for ancestral graphs. The Annals of Statistics 37(5B):2808–2837

Banerjee O, El Ghaoui L, d'Aspremont A (2008) Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. Journal of Machine Learning Research 9:485–516

Chow CK, Liu CN (1968) Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory 14:462–467

Claeskens G, Hjort N (2003) The focused information criterion. Journal of the American Statistical Association 98:900–916, with discussion and a rejoinder by the authors

Claeskens G, Hjort N (2008a) Minimising average risk in regression models. Econometric Theory 24:493–527

Claeskens G, Hjort N (2008b) Model Selection and Model Averaging. Cambridge University Press, Cambridge

Cox DR, Wermuth N (1996) Multivariate dependencies: Models, Analysis and Interpretation. Chapman & Hall, London

Dempster A (1972) Covariance selection. Biometrics 28(1):157–175

Dor D, Tarsi M (1992) A simple algorithm to construct a consistent extension of a partially oriented graph. Tech. rep.

Drton M, Perlman M (2004) Model selection for Gaussian concentration graphs. Biometrika 91(3):591–602

Drton M, Perlman M (2008) A SINful approach to Gaussian graphical model selection. Journal of Statistical Planning and Inference 138(4):1179–1200

Drton M, Richardson T (2004) Iterative conditional fitting for Gaussian ancestral graph models. In: Chickering D, Halpern J (eds) Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pp 130–137

Edwards D (2000) Introduction to Graphical Modelling, 2nd edn. Springer, New York

Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3):432–441

Friedman JH (1991) Multivariate adaptive regression splines. The Annals of Statistics 19(1):1–67

Gammelgaard Bøttcher S (2004) Learning bayesian networks with mixed variables. PhD thesis, Aalborg University

Grossman D, Domingos P (2004) Learning bayesian network classifiers by maximizing conditional likelihood. In: Brodley C (ed) Proceedings of the 21st International Conference on Machine Learning

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning, 2nd edn. Springer Series in Statistics, Springer, New York, data mining, inference, and prediction

Heckerman D, Geiger D (1995) Learning bayesian networks: A unification for discrete and gaussian domains. In: Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, pp 274–284

Hjort N, Claeskens G (2003) Frequentist model average estimators. Journal of the American Statistical Association 98:879–899, with discussion and a rejoinder by the authors

Hjort N, Claeskens G (2006) Focussed information criteria and model averaging for Cox's hazard regression model. Journal of the American Statistical Association 101:1449–1464

Hjort NL (1994) The exact amount of t-ness that the normal model can tolerate. Journal of the American Statistical Association 89:665–675

Jardine N, van Rijsbergen C (1971) The use of hierarchic clustering in information retrieval. Information Storage and Retrieval 7(5):217–240

Kalisch M, Mächler M, Colombo D, Maathuis M, Bühlmann P (2012) Causal inference using graphical models with the R package pcalg. Journal of Statistical Software 47(11):1–26

Koller D, Friedman N (2009) Probabilistic Graphical Models: Principles and Techniques. MIT Press, Cambridge, MA

Krishnamurthy V, Ahipaşaoğlu S, d'Aspremont A (2012) A pathwise algorithm for covariance selection. In: Sra S, Nowozin S, Wright S (eds) Optimization for Machine Learning, MIT Press, pp 479–494

Lauritzen S (1996) Graphical Models. Oxford University Press

Li L, Toh KC (2010) An inexact interior point method for l_1-regularized sparse covariance selection. Mathematical Programming Computation 2(3-4):291–315

Mansour J, Schwarz R (2008) Molecular mechanisms for individualized cancer care. Journal of the American College of Surgeons 207(2):250 – 258

Mardia KV, Kent JT, Bibby JM (1979) Multivariate Analysis. Academic Press, London

Meinshausen N, Bühlmann P (2006) High-dimensional graphs and variable selection with the lasso. The Annals of Statistics 34(3):1436–1462

Pearl J (1988) Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc.

Richardson T, Spirtes P (2002) Ancestral graph Markov models. The Annals of Statistics 30(4):962–1030

Schmidt M, Niculescu-Mizil A, Murphy K (2007) Learning graphical model structure using l_1-regularization paths. In: Proceedings of the 22nd AAAI Conference on Artificial Intelligence, AAAI Press, pp 1278–1283

Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6(2):461–464

Scutari M (2010) Learning bayesian networks with the bnlearn R package. Journal of Statistical Software 35(3):1–22

Shastry BS (2006) Pharmacogenetics and the concept of individualized medicine. The Pharmacogenomics Journal 6(1):16–21

Spirtes P, Meek C, Richardson T (1999) An algorithm for causal inference in the presence of latent variables and selection bias. In: Glymour C, Cooper G (eds) Computation, Causation and Discovery, MIT Press, pp 211–252

Spirtes P, Glymour C, Scheines R (2000) Causation, Prediction and Search, 2nd edn. MIT Press, Cambridge, MA

Tsamardinos I, Brown EL, Aliferis CF (2006) The max-min hill-climbing Bayesian network structure learning algorithm. Journal of Machine Learning Research 65(1):31 – 78

van 't Veer L, Bernards R (2008) Enabling personalized cancer medicine through analysis of gene-expression patterns. Nature 452(7187):564–570

Whittaker J (1990) Graphical Models in Applied Multivariate Statistics. John Wiley & Sons, Chichester

Williamson J (2005) Bayesian Nets and Causality. Philosophical and Computational Foundations. Oxford University Press, Oxford

Witten DM, Friedman JH, Simon N (2011) New insights and faster computations for the graphical lasso. Journal of Computational and Graphical Statistics 20(4):892–900

Yuan M, Lin Y (2007) Model selection and estimation in the Gaussian graphical model. Biometrika 94(1):19–35

Zhang J (2008) On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artificial Intelligence 172(16):1873 – 1896

Zhang X, Liang H (2011) Focused information criterion and model averaging for generalized additive partial linear models. The Annals of Statistics 39(1):174–200

Zhao T, Liu H, Roeder K, Lafferty J, Wasserman L (2012) The huge package for high-dimensional undirected graph estimation in R. Journal of Machine Learning Research 13:1059–1062